Department of Social Systems and Management

Discussion Paper Series

No.1228

Transpersonal Understanding through Social Roles, and Emergence of Cooperation

By

Mamoru KANEKO and J. Jude KLINE

March 2009

UNIVERSITY OF TSUKUBA
Tsukuba, Ibaraki 305-8573
JAPAN

Transpersonal Understanding through Social Roles, and Emergence of Cooperation* †

Mamoru Kaneko‡and J. Jude Kline §

14 April 2009

Abstract

Inductive game theory has been developed to explore the origin of beliefs of a person from his accumulated experiences of a game situation. So far, the theory has been restricted to a person's view of the structure not including another person's thoughts. In this paper, we explore the experiential origin of one's view of the other's beliefs about the game situation. We restrict our exploration to a 2-role (strategic) game, which has been recurrently played by two people who occassionally switch roles. By switching roles, each person accumulates experiences of both roles and these experiences become the source of his transpersonal view about the other. Reciprocity in the sense of role-switching is crucial for deriving his own and the other's beliefs. We consider how a person can use these for his behavior revision, and we define an equilibrium called an intrapersonal coordination equilibrium. Based on this concept, we show that cooperation will emerge as the degree of reciprocity increases.

1. Introduction

We will consider the problem of how a person obtains beliefs[1] about other persons' thoughts. We look for experiential bases for such beliefs. A crucial distinction is made

*The authors are partially supported by Grant-in-Aids for Scientific Research No.17653018, Ministry of Education, Science and Culture, and Australian Research Council Discovery Grant DP0560034.

†The authors thank Nathan Berg, Burkhard Schipper, Bill Schworm, and participants in lectures given at Waseda University for valuable discussions on the subject of the present paper. They also thank Yusuke Narita for detailed comments on an earlier version of the paper.

‡Institute of Policy and Planning Sciences, University of Tsukuba, Ibaraki 305-8573, Japan (kaneko@shako.sk.tsukuba.ac.jp)

§Department of Economics, School of Business, Bond University, Gold Coast, QLD 4229, Australia, (Jeffrey_Kline@bond.edu.au)

[1]We use the term "belief" and allow it to include "knowledge". Our "belief" is about a structure but not probabilistic. Here, we require some justification in the form of evidence for beliefs about the other.

between persons (actors) and social roles (players), which allows a person to switch roles from time to time. This enables a person, based on his experiences, to guess the other person's thinking, and even to obtain a social perspective, which goes beyond an individual perspective. Within this framework, we can go further to discuss the emergence of cooperation.

In this introduction, we will refer to the standard game theory and relevant literatures so as to better understand our approach. Then we discuss new concepts to be needed and phenomena to be captured in the scope of our approach.

1.1. General Motivations

It is customary in game theory and economics to assume well-formed beliefs of a game for each player, which is often implicit and sometimes explicit. The present authors [16], [17] and [18] have developed inductive game theory in order to explore the basic question of where a personal understanding of a game comes from[2]. In those papers, an individual view and its derivation from a player's experiences are discussed from various points of view. Nevertheless, they did not reach the stage of research on his thoughts about other persons' thoughts. This paper aims to take one step further to explore the origin (and emergence) of a person's thoughts about other persons' thoughts.

To take this step, a person needs to think about others' beliefs on the social structure. We introduce the concept of *social roles,* and use also the term, *person*, to distinguish it from the standard term "*player*"; the latter is close to our notion of a social role. A person takes a social role (exogenously given), and may switch his role from time to time. Taking different roles will be a key to understanding others' perspectives. By projecting his experiences of the various roles in his mind, he develops his social perspective including others' thoughts. In the following, we confine ourselves to the 2-person case to focus on the main problems emerging from those new concepts.

When the persons switch social roles reciprocally, a new feature is emerging: Reciprocal relationships provide each person with a rich source for inferring/guessing the beliefs of the other person[3]. When the persons switch roles enough, each has been in the same position and has seen the other person in the corresponding position. This level of reciprocity may give each person "reason to believe" that the other's view is the same as his. This idea is reminiscent of a requirement imposed for "common knowledge" in Lewis [21], which is more similar to the fixed-point characterization of "common knowl-

[2]A seminal form of inductive game theory was given in Kaneko-Matsui [19]. An alternative formulation was given in Matsui [22].

[3]We use the term "reciprocal" in the sense of "performed, experienced, or felt by both sides" as (3) of the American Heritage dictionary (1980). It is used in the evolutionary and behavioral game theory literature to mean a type of "tit for tat" behavior (see Camerer [3] and Gardenfors [8] for such alternative uses).

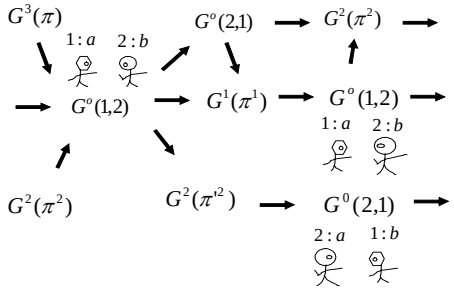

Figure 1.1: Social Web

edge" than the infinite hierarchy of knowledge (cf., Fagin-Halpern-Moses-Vardi [7] and Kaneko [14]). This will be the key for the development of our theory.

Broadly speaking, we may regard our exploration as undertaken along the line of *symbolic interactionism* due to Mead [23] (cf., Collins [5], Chap.7). Each isolated experience is not more than a sequence or a set of symbols. However, by playing roles reciprocally and interactively, the accumulated set of experiences could constitute some meaning. This is analogous also to symbolic logic (cf., Mendelson [24] and Kaneko [14]) in that it starts with primitive symbols without meanings. Formulae consisting of those symbols and their further combinations may eventually generate some meanings. An individual perspective is obtained by combining experiences as sequences of meaningless symbols. A social perspective is obtained by combining experiences of reciprocal interactions into an even greater view. In the sociology literature, these problems were discussed without giving a mathematical formulation. Our approach is regarded as a mathematical formulation of symbolic interactionism, and expands its perspective while enabling us to examine it critically.

An example, due to Mead [23], for the distinction between a person and a social role consists of the positions of pitcher, catcher, first base, etc..., in a baseball team. Since we use only a 2-role strategic game for our exploration, it may be better to refer to a 2-role example of a family affair between a wife and a husband: They may divide their housekeeping into the breakfast maker and dinner maker. There are numerous alternative varieties, e.g., raising children versus working at the office, cleaning the house versus gardening, or allocating finances versus generating finances. In such situations, role-switching becomes crucial for understanding the other's perspective.

A target game situation is in a social web like Fig.1.1: Two persons 1 and 2 play the strategic game $G^o(1,2)$ in the north-west in Fig.1.1, where G^o is assumed to be a standard strategic game with two "players", which are roles a and b. In $G^o(1,2)$,

persons 1 and 2 take roles a and b, respectively. If they switch roles a and b, the game situation becomes as $G^o(2,1)$ in the south-east. Although we will focus on a particular situation such as G^o, it is a small part of the entire social web for the persons. Each person participates in various other social games such as university administration, a community baseball team, etc. This remark should not be forgotten, and will be discussed in various places in the paper.

It is a salient point of our approach that thinking about the other's thoughts in one's mind might lead to cooperation. This type of idea was discussed and emphasized by Mead [23] and his predecessor, Cooley [6] to argue the pervasiveness of cooperation in human society. This level of optimism was criticized as too naive by later sociologists (see Collins [5], Chap.7). In our theory, cooperation is one possibility, but not necessarily guaranteed. We can discuss when cooperation likely happens and when not.

It is another salient point that the inductive game theory approach, especially the development in this paper, gives some answers to many of the "Top Ten Research Questions" given in Camerer [3], e.g., "*How do people value the payoffs of others?*", and "*What game do people think they are playing?*" We will address these questions.

1.2. Basic Postulates for an Understanding of the Other's Mind

Kaneko-Kline [16], [17] and [18] chose a general environment corresponding to an extensive game and already met a lot of basic problems in the consideration of experiences and their generations. Also, various basic notions in the extant game theory such as "information", "memory", and moreover, "extensive game" itself needed to be redefined. In the consideration of induction, they met also partiality, indeterminacy, falsity, etc. in an inductively derived view on the social structure.

As stated above, we confine ourselves here to a 2-role strategic game to avoid the difficulties mentioned above. Nevertheless, we now include the other's thoughts: We need various subtle definitions. Thus, it would be better to mention the basic (premathematical) postulates for one's thinking about the other's and for the emergence of cooperation[4].

First, we make the *basic postulate* that a person cannot directly look into the other's mind. Instead, we postulate that person i infers/guesses from his own experiences what person j may know about the situation. Transpersonal projection of one's experiences onto the other is considered based on experiences of different roles. Thus, our theory is experiential and follows the tradition from Mead [23].

If person 1 has experienced two roles a and b from time to time, person 1 could infers/guesses person 2's experiences and thoughts. This requires reciprocity of roles played by those two persons. We will explore how such reciprocity is needed for person

[4]In our premathematical arguments, we use the term "postulate". This term means simply a starting assumption to facilitate our discourse.

1 to fully imagine the other's thoughts. Another extreme case, which should not be ignored, is one where they do not switch roles at all, and as a consequence, person 1 cannot imagine 2's thoughts. Our theory presents some capacity to separate these cases and generates different results based on this separation. One such difference is that cooperation would not be reached without a sufficient level of reciprocity.

One more postulate we should mention here is on use of the beliefs about the other's thoughts. With role-switching, a person can begin to think about a change in his behavior and of how the other thinks of this change. Incorporating his transpersonal projection of one's experiences onto the other's thoughts, we define the equilibrium concept called an *intrapersonal coordination equilibrium*. Our analysis of the emergence of cooperation is based on this concept.

1.3. Brief Discussions on Cooperative Behavior in the Literature

In the game theory literature, cooperative behavior has been extensively discussed, but no relationships between cooperative behavior and cognitive assumptions are discussed. Since this will be important to distinguish our new theory from the other extant theories, we will give brief discussions on the treatments of cooperative behavior in the game theory literature. Here, we will look only at cooperative game theory, the Nash program, and the repeated game approach.

Cooperative game theory was already extensively discussed in von Neumann- Morgenstern [28] and a lot of branches have been developed. In them, cooperation itself is a very basic postulate, and possible outcomes resulting from cooperative behavior are targets to be studied. This theory does not address the question of the origin of cooperation and is incapable in discussing this question.

The *Nash program*, which was originally suggested by Nash [25], p.295, may appear to resolve the incapability by reducing cooperation into individual activities for cooperation: The possibilities for some players to propose to cooperate with some other players are described as moves in (rules of) an extensive game. In this theory, we may discuss a process for cooperative behavior, an example of which was given in Nash [26]. The Nash program reduces the postulate of cooperation into the rules of a game, but this theory does not address the question of emergence of cooperation.

The *repeated game approach* (cf., Hart [11]) has two similar aspects to our approach in that both treat recurrent situations and discuss cooperation as a possible outcome. Nevertheless, the two approaches have a radical difference in their basic cognitive postulates. Also, in the repeated game approach, a *cooperative outcome* is based on threats but not on behavior to cooperate with the other. In our theory, cooperative behavior becomes possible with the cognition of the other's beliefs when the degree of reciprocity increases.

Rigorously speaking, the repeated game approach formulates the entire situation as

a huge one-shot game, i.e., an infinite extensive game. Then the Nash equilibrium (or its refinement) is adopted for this entire game. The Nash equilibrium is interpreted as describing *ex ante* decision making in the sense that each player makes a decision as well as his prediction about the others' decision before the actual play of the repeated game. This requires each player to be fully cognizant of the entire game structure[5]. For this reason, the repeated game approach cannot address the basic cognitive question of where beliefs about the game structure and others' beliefs comes from for a player. In this respect, the Nash program is in the same position.

Thus, the extant theories do not address the question of origin of beliefs for players, and furthermore, their postulates are not suitable to a study of an emergence of cooperation. This should not be taken to mean that cooperation does not prevail in society. Contrary to this, it is believed among many social scientists, as stated in Section 1.1, that cooperation and cooperative behavior are widely observed phenomena in society. Inductive game theory can discuss both the emergence of beliefs and cooperation.

We will connect our cooperation result to some behavioral game theory literature. Behavioral/experimental game theory has reported many experiments to support the pervasiveness of cooperative behavior. One observation in the repeated situation of the prisoner's dilemma is that cooperative outcomes emerge after some repetition of the game (cf., Cooper-DeJong-Forsyth-Ross [4]) Another is the experimental study of the ultimatum game and dictator game, which shows that people do cooperate, even though the standard game theoretical argument (subgame perfection) does not predict cooperation at all (cf., Güth-Schmittberger-Schwarze [9], Kahneman-Knetsch-Thaler [13], and also Camerer [3] for a more recent survey). In Section 7, we will examine implications of our theory of cooperation to the literature of those behavioral studies, specifically, looking at the prisoner's dilemma, ultimatum game and dictator game.

The remainder of the paper is as follows. Section 2 gives the basic definitions of a 2-role game, the domain of experiences, etc. Section 3 defines person's *direct understanding* of the basic situation and *transpersonal understanding* of the other's understanding from his experiences, which is an intermediate step to the main definition of an inductively derived view (i.d.view) given in Section 4. The i.d.view combines those understandings together with the *regular behavior* and *frequency weights* of roles. In Section 5, the definition of an intrapersonal coordination equilibrium is defined, and is studied, first, in non-reciprocal cases. In Section 6, we study it in reciprocal cases. The results obtained Sections 5 and 6 are applied to the prisoner's dilemma, ultimatum game and dictator game in Section 7. In Section 8, we will discuss external and reciprocal relations between the persons. In Section 9, we will discuss implications of our approach together with the results obtained in this paper.

[5]This is not the intended interpretation of a Nash equilibrium in the repeated game for some authors (e.g., Axelrod [2]) - - in which case, the cognitive assumption must be different from the full cognizance but has not been explicated.

2. Two-Person Strategic Game with Social Roles

2.1. 2-Role Strategic Game and Role Assignments

We start with a 2-*role (strategic) game* $G = (a, b, S_a, S_b, h_a, h_b)$, where a and b are (social) *roles*, $S_r = \{s_{r1}, ..., s_{r\ell_r}\}$ is a finite set of *actions,* and $h_r : S_a \times S_b \to \mathbf{R}$ is a *payoff function* for each role $r = a, b$. We will refer to this game as the *base game.* Each role is taken by *person* $i = 1, 2$. We have a *role assignment* π, which is a one-one mapping $\pi : \{a, b\} \to \{1, 2\}$. The expression $\pi(r) = i$ means that i is the person assigned to role r. We may also write $\pi = (i_a, i_b)$ to mean that persons i_a and i_b take roles a and b, respectively.

A 2-*person (strategic) game with social roles* is given by adding a role assignment $\pi = (i_a, i_b)$ to a 2-role strategic game G:

$$G(\pi) = (i_a, i_b, S_a, S_b, h_a, h_b). \tag{2.1}$$

That is, persons i_a and i_b taking roles a and b play the base game G. We consider the following example, which will be used later.

Example 2.1: In the game $G(1, 2)$ of Table 2.1, persons 1 and 2 are assigned to roles a and b. The game $G(2, 1)$ has the same structure, but the role-assignments are reversed. A larger recurrent social context exists behind games $G(1, 2)$ or $G(2, 1)$, like Fig.1.1. In Fig.1.1, $G^0(1, 2)$ and $G^0(2, 1)$ are two local situations with the same 2-role game G^0. We assume that the persons behave in a regular manner subject to some trial deviations and that each person accumulates experiences of playing this game with different roles.

Table 2.1; $G(1, 2)$

$1\backslash 2$	$\mathbf{s}_{b1}$	$\mathbf{s}_{b2}$	$\mathbf{s}_{b3}$
$\mathbf{s}_{a1}$	$(3, 3)$	$(10, 2)$	$(3, 1)$
$\mathbf{s}_{a2}$	$(2, 10)$	$(4, 4)$	$(5, 5)$
$\mathbf{s}_{a3}$	$(1, 3)$	$(5, 5)$	$(4, 4)$

Since the situation we consider is recurrent, the information structure of observations after each play of a game should be specified. We assume that after each play of $G(\pi)$, each person with role $\pi(r) = i$ observes

Ob1: the action pair (s_a, s_b) played;

Ob2: his own payoff (value) from this pair.

These postulates are asymmetric in that person i can observe both actions taken by him and the other, but can observe only his own payoff. This asymmetry will be important in Section 3. With respect to the treatment of payoffs, we should emphasize the distinction between *having* a payoff function and *knowing* it. Here, we assume that each person

recognizes each payoff value $h_r(s_a, s_b)$ only when he experiences it but does not know the function h_r itself. Only after he has accumulated enough memories of experiences, he may come to know some part of the payoff function.

2.2. Accumulated Memories

Now, we consider person i's accumulation of experiences up to a particular point of time. It is summarized as a *memory kit* $\kappa_i = \langle (s_a^o, s_b^o), (D_{ia}, D_{ib}), (h_{ia}, h_{ib}); (\rho_{ia}, \rho_{ib}) \rangle$, which consists of

$\kappa 1$: the pair (s_a^o, s_b^o) of *regular actions*;

$\kappa 2$: the *accumulated domain of experiences* $D_i = (D_{ia}, D_{ib})$ consisting of experiences of action pairs from taking roles a and b, respectively;

$\kappa 3$: person i's *observed payoff functions* (h_{ia}, h_{ib}) over D_i;

$\kappa 4$: person i's vector (ρ_{ia}, ρ_{ib}) of (subjective) *frequency weights* for roles a and b.

Person i has obtained these components by playing game G with possibly different roles from time to time. Component $\kappa 1$ means that the persons play regularly the actions s_a^o and s_b^o when they are assigned to roles a and b. Component $\kappa 2$ states that person i has other experiences in addition to the regular actions. Occasionally, each person i deviates from s_r^o to some other actions s_r, and some (or all) actions experienced are remaining in his mind, which form the sets D_{ia} and D_{ib}. The third components, (h_{ia}, h_{ib}), in $\kappa 3$ are the observed (perceived) payoff functions over (D_{ia}, D_{ib}), which are mathematically defined presently. The last component (ρ_{ia}, ρ_{ib}) in $\kappa 4$ means that person i evaluates subjectively how frequently he has been assigned to roles a and b. Accurate weights are not really our intention[6], but here we assume that it is a single vector for each i.

In the following, we use the convention that if $r = a$ or $r = b$, then $s_{(-r)} \equiv s_{-r} = s_b$ or s_a, respectively, but $(s_r; s_{-r}) = (s_a, s_b)$ in either case.

Mathematically, the components of a memory kit $\kappa_i = \langle (s_a^o, s_b^o), (D_{ia}, D_{ib}), (h_{ia}, h_{ib}); (\rho_{ia}, \rho_{ib}) \rangle$ are given and assumed to satisfy the following conditions: for all $r = a, b$ and $s_r \in S_r$:

$$(s_a^o, s_b^o) \in D_{ia} \cup D_{ib} \subseteq S_a \times S_b; \tag{2.2}$$

$$\text{if } (s_a, s_b) \in D_{ir}, \text{ then } (s_a, s_b^o) \in D_{ir} \text{ and } (s_a^o, s_b) \in D_{ir}; \tag{2.3}$$

$$h_{ir} : D_{ir} \to \mathbf{R} \text{ and } h_{ir}(s_a, s_b) = h_r(s_a, s_b) \text{ for all } (s_a, s_b) \in D_{ir}; \tag{2.4}$$

$$\rho_{ia} + \rho_{ib} = 1 \text{ and } \rho_{ia}, \rho_{ib} \geq 0; \tag{2.5}$$

$$\text{if } \rho_{ir} = 0, \text{ then } D_{ir} = \emptyset. \tag{2.6}$$

[6]See Hu [12] for the concept of frequency and the frequentist interpretation of expected utility theory.

Condition (2.2) states that the domains of accumulation include the regular actions (s_a^o, s_b^o). It is the intent that (s_a^o, s_b^o) has been played in $G(1,2)$ and $G(2,1)$ as the regular actions, while person i has made some trial deviations from (s_a^o, s_b^o) and accumulated his experiences in D_{ia} and D_{ib}. We allow D_{ia} or D_{ib} to be empty, though the union $D_{ia} \cup D_{ib}$ is nonempty by (2.2). If $D_{ia} = \emptyset$, then person i has never experienced role a at least in his memory.

Condition (2.3) states that if ever some pair (s_a, s_b) is accumulated in D_{ir}, then the pairs (s_a, s_b^o) and (s_a^o, s_b) coming from the unilateral trials of s_a and s_b from the regular actions (s_a^o, s_b^o) are also accumulated. It expresses the idea that the domain of accumulation is generated by unilateral trials from the regular action. This will be briefly discussed in Section 2.3.

Condition (2.4) states that person i knows a functional relationship between each pair $(s_a, s_b) \in D_{ir}$ and the payoff value from it when he takes role r. To avoid confusions with the objective payoff function h_r, we define the function $h_{ir} : D_{ir} \to \mathbf{R}$. Thus, this is the experienced payoff function of person i when he takes role r. Mathematically, h_{ir} is the restriction of h_r to D_{ir}. Condition (2.5) states that (ρ_{ia}, ρ_{ib}) is a vector of subjective frequency weights. We do not require that these subjective frequency weights are precisely the same as the objective frequencies. For example, $\rho_{ir} = 0$ is interpreted as person i has no recollection of being in role r, even if he was there objectively with some negligible frequency. Similarly, $\rho_{ir} = 1/2$ is interpreted as meaning in person i's mind, the switching is reciprocal, while objectively, the frequency may be slightly different from $1/2$.

In keeping with the subjective interpretation, (2.6) states that if person i has $\rho_{ir} = 0$, then he has no recollection of being in that role.

The following lemma is proved by using (2.3) twice. It states that if person i has some experience at role r in his mind, then he accumulated at least the pair of regular actions at that role.

Lemma 2.1. If $D_{ir} \neq \emptyset$, then $(s_a^o, s_b^o) \in D_{ir}$.

We use the following terms: When $(s_r; s_{-r}^o) \in D_{ir}$, it is called an *active experience* (*deviation*) for person i at role r; and when $(s_r; s_{-r}^o) \in D_{i(-r)}$, it is a *passive experience* for person i at role $-r$. That is, if one person makes a deviation, and if it remains in his domain, it is an active experience, and if it remains in the domain of the other person, it is a passive experience for that person.

In this paper, reciprocity plays an important role, but we have various notions of and degree of reciprocities. One important reciprocity is between the domains D_{ia} and D_{ib} for a fixed person i. We will have a strong form of reciprocity over those domains when there is a sufficient amount of reciprocity in role-switching. We say that the domains (D_{ia}, D_{ib}) for person i is *strongly internally reciprocal* iff

$$D_{ia} = D_{ib}. \tag{2.7}$$

It involves a comparison only of person i's domains D_{ia} and D_{ib}.

In fact, (2.7) is stronger than what we will target in this paper. For the weaker version, first we define the set $\text{Proj}(T) := \{(s_a, s_b) \in T : s_a = s_a^o \text{ or } s_b = s_b^o\}$. Then, (2.7) is weakened to

$$\text{Proj}(D_{ia}) = \text{Proj}(D_{ib}), \tag{2.8}$$

in which case, we say that D_{ia} and D_{ib} are *internally reciprocal.* This requires the equivalence of these sets up to only unilateral changes from the regular actions (s_a^o, s_b^o).

We should bear in mind that since the experiences in $D_{ia} \cup D_{ib}$ are generated both by person i and another person j, some external reciprocal relationships between i and j are the background for condition (2.8) or (2.7). However, we will focus first on person i's internal thoughts such as inferences/guesses from his own experiences, so we postpone our discussions about the background external relationships until Section 8.

Let us consider several examples for the domains (D_{1a}, D_{1b}) and (D_{2a}, D_{2b}). In the following examples, we assume for simplicity that each person makes trials with all actions at the role he has assigned to.

(1)(Non-reciprocal Domains): In these domains, the persons do not switch the roles at all. First, we consider the *non-reciprocal active domains.* Let $D_1^N = (D_{1a}^N, D_{1b}^N)$ and $D_2^N = (D_{2a}^N, D_{2b}^N)$ be given as follows:

$$\begin{aligned} D_{1a}^N &= \{(s_a, s_b^o) : s_a \in S_a\} \text{ and } D_{1b}^N = \emptyset \\ D_{2a}^N &= \emptyset \text{ and } D_{2b}^N = \{(s_a^o, s_b) : s_b \in S_b\}. \end{aligned} \tag{2.9}$$

With these domains, neither (2.7) nor (2.8) holds. Each person makes deviations over all his actions. However, each accumulates only active experiences, which means that he is either insensitive to (or ignores) the deviations by the other person. In this example, it is natural to assume that $\rho_{1a} = \rho_{2b} = 1$.

We mention that there are other non-reciprocal domains. For example, the *non-reciprocal active-passive domain* $D_{1a}^{NAP} = D_{1a}^N \cup \{(s_a^o, s_b) : s_b \in S_b\}$ and $D_{1b}^{NAP} = \emptyset$ describes the non-reciprocal case where person 1 is sensitive to both active and passive deviations. It is defined similarly for person 2. They are not yet internally reciprocal, while each person is sensitive to the other's trials.

We have numerous varieties of reciprocal domains where the roles are switched. We focus on two reciprocal cases in particular.

(2):(Reciprocal Active Domain): The *reciprocal active domain* $D_1^A = (D_{1a}^A, D_{1b}^A)$ for person 1 is given as:

$$D_{1a}^A = \{(s_a, s_b^o) : s_a \in S_a\} \text{ and } D_{1b}^A = \{(s_a^o, s_b) : s_b \in S_b\}. \tag{2.10}$$

This means that person 1 makes trials with all actions for each role $r = a, b$, but he is insensitive to person 2's trials. If person 2 behaves in the same manner, then $D_{2a}^A = D_{1a}^A$

and $D^A_{2b} = D^A_{1b}$. Although both persons' domains are the same, the internal reciprocity condition (2.8) does not hold.

We give one domain that is internally reciprocal.

(3)(Reciprocal Active-Passive Domain): The *reciprocal active-passive domain* $D^{AP}_1 = (D^{AP}_{1a}, D^{AP}_{1b})$ is given as:

$$D^{AP}_{1a} = D^{AP}_{1b} = \{(s_a, s^o_b) : s_a \in S_a\} \cup \{(s^o_a, s_b) : s_b \in S_b\}. \tag{2.11}$$

Person 1 makes trials with all actions across both roles, and he is sensitive to both active and passive "unilateral" trials, but not joint-trials.[7] If person 2 has the same personality, then 2 has the same domains: $D^{AP}_{2a} = D^{AP}_{1a}$ and $D^{AP}_{2b} = D^{AP}_{1b}$. This domain satisfies (2.8) and even (2.7), and is still smaller than the full reciprocal domain defined by $D^F_{ir} = S_a \times S_b$ for $i = 1, 2$, and $r = a, b$.

2.3. An Informal Theory of Behavior and Accumulation of Memories

Our mathematical theory starts with a memory kit. Behind a memory kit, there is some underlying process of behavior and accumulation of memories. We now describe one such underlying process informally. Some parts of the following informal theory are more precisely discussed for the one-person case in Akiyama-Ishikawa-Kaneko-Kline [1].

(1): Postulates for Behavior and Trials: In the recurrent situation, the role-switching is given exogenously, and we do not consider endogenous efforts for role-switching. We state this as a postulate.

Postulate BH0 (Switching the Roles): The role assignment changes from time to time, which is exogenously given.

The next postulate is the rule-governed behavior of each person in the recurrent situation $..., G^o(1,2), G^o(2,1), ..., G^o(1,2), ...$

Postulate BH1 (Regular actions): Each person typically behaves following the regular action s^o_r when he is assigned to role r.

It may be the case that the regular actions are person-dependent, but in this paper, we simply assume that both persons follow the same regular action for each role. Person i may have adopted the regular actions s^o_a and s^o_b for roles a and b for some time without thinking, perhaps since he found it worked well in the past or he was taught to follow it. Without assuming regular actions and/or patterns, a person may not be able to extract any causality from his experiences. In essence, learning requires some regularity.

To learn some other part than the regular actions, the persons need to make some trial deviations. We postulate that such deviations take place in the following manner.

[7] One reason could be that joint trials are too infrequent, and his sensitivity is not strong enough to recall them.

Postulate BH2 (Occasional Deviations): Once in a while (infrequently), each person, taking role r, unilaterally and independently makes a trial deviation $s_r \in S_r$ from his regular action s_r^o, and then returns to his regular action s_r^o or s_{-r}^o.

Early on, such deviations may be unconscious and/or not well thought out. Nevertheless, a person might find that a deviation leads to a better outcome, and he may start making deviations consciously. Once he has become conscious of his behavior-deviation, he might make more and/or different trials.

Postulate BH2 justifies condition (2.3) since it implies that only one person's deviation more likely occurs than both persons'.

(2): Cognitive Postulates: Each person may learn something through his regular actions and deviations. What he learns in an instant is described by his local (short-term) memory. It takes the form of $\langle r, (s_a, s_b), h_{ir}(s_a, s_b) = h_r(s_a, s_b)\rangle$. Once this triple is transformed to a *long-term memory*, D_{ir} is extended into

$$D_{ir} \cup \{(s_a, s_b)\},$$

and "$h_{ir}(s_a, s_b) = h_r(s_a, s_b)$" is also recorded in the memory kit κ_i, which is given in (2.4). For the transition from local memories to long-term memories, there are various possibilities. Here we list some postulates based on bounded memory abilities.

The first states that if a short-term memory does not occur frequently enough, it will disappear from the mind of a person. We give this as a postulate for a cognitive bound on a person.

Postulate EP1 (Forgetfulness): If experiences are not frequent enough, then they would not be transformed into a long-term memory and disappear from a person's mind.

This is a rationale for not assuming that a person has a full record of local memories. If it is not reinforced by other occurrences or the person is very conscious, they may disappear from his mind.

In the face of such a cognitive bound, only some memories become lasting. The first type of such memories are the regular ones since they occur quite frequently. The process of making a memory last by repetition is known as habituation.

Postulate EP2 (Habituation): A local (short-term) memory becomes lasting as a long-term memory in the mind of a person by habituation, i.e., if he experiences something frequently enough, it remains in his memory as a long-term memory even without conscious effort.

By EP2, when the persons follow their regular actions, the local memories given by them will become long-term memories by habituation.

A pair obtained by only one person's deviation remains next likely, which supports (2.3). We postulate that a person may consciously spend some effort to memorize the outcomes of his own trials.

Postulate EP3 (Conscious Memorization Effort): A person makes a conscious effort to memorize the result of his own trials. These efforts are successful if they occur frequently enough relative to his trials.

In this paper, we will sometimes make use of a postulate for a different degree of sensitivity for active and passive experiences.

Postulate EP4 (Sensitive with Active relative to Passive): A person is more (or not less) sensitive to his own active deviation than he is to his passive experiences.

We adopt this postulate as a starting point. It may need empirical tests to determine which forms are more prominent in society. In this paper, however, we will simply take the *relativistic attitude* that a person's domain is not uniquely determined but takes various possible forms.

3. Direct and Transpersonal Understandings from Experiences

When a person considers the situation described by the 2-role strategic game G based on his accumulated experiences, he meets two problems: (1) his own understanding about G; and (2) his understanding of the other's thoughts about G. The former is straightforward in that it simply combines his experiences, while the latter needs some additional interpersonal thinking. In this section, we describe how a person might deal with these two problems. We do not yet include the regular actions (s_a^o, s_b^o) and frequency weights (ρ_{ia}, ρ_{ib}), which will be taken into account in the definition of an inductively derived view to be given in Section 4.

3.1. Transpersonal Postulates for the Other's Thoughts

First, we state our basic ideas on how a person deals with the above mentioned problems as postulates. We adopt experientialism for these postulates. The first postulate is about a person's direct understanding of a situation, which refers to the problem (1).

Postulate DU1 (Direct Understanding of the Object Situation): A person combines his accumulated experiences to construct his view on the situation in question.

This will be presently formulated as a direct understanding g^{ii}.

Now, consider how a person thinks about the other's understanding. We adopt two new postulates for it, which we call *transpersonal postulates*. A metaphor may help the reader understand those postulates:

$*1$ *The agony of a broken heart can only be understood by a person whose heart was once broken;*

$*2$ *yet, he doubts her agony because he cannot explain her broken heart.*

The part $*1$ corresponds to the following postulate:

Postulate TP1 (Projection of Self to the Other): A person projects his own experienced payoff onto the other person if he believes that the other knows his payoff at that experience.

By postulate Ob2, he observes only his own payoff. To think about the other's payoff, he uses also his own experienced payoff. By postulate TP1, we propose that a person projects his own experiences onto the other. We could use an alterative postulate, e.g., I find by experience that you are different from me; this however, happens rarely. A person keeps TP1 as his principle until he finds enough counter evidence. We regard projection of oneself as a very basic postulate.

Notice that postulate TP1 is a conditional statement. We require some evidence for a person to believe that the other knows the payoff, which is expressed as the next postulate. It corresponds to $*2$ of the above metaphor.

Postulate TP2 (Experiential Reason to Believe): A person believes that the other knows a payoff only when the person has a sufficient experiential reason for the other to have the payoff.

In the above metaphor, having a broken heart is an experience of losing a love, and it causes agony. Postulate TP1 requires that the agony caused by losing a love is understood by projecting one's past experience, which is $*1$. Then, postulate TP2 requires some experiential evidence (reason) to believe that she has broken heart. This is expressed as its contrapositive in $*2$: Since he has no experiential reason to believe her broken heart, he doubts her agony. This "reason to believe" is reminiscent of a requirement for the concept of "common knowledge" in Lewis [21]. In the next section, we will give an explicit formulation of the other's understanding based on postulates TP1 and TP2.

3.2. Direct and Transpersonal Understandings

Suppose that person i has accumulated his experiences in a memory kit $\kappa_i = \langle (s_a^o, s_b^o), (D_{ia}, D_{ib}), (h_{ia}, h_{ib}); (\rho_{ia}, \rho_{ib}) \rangle$. He, now, constructs his *direct understanding* of the game situation including own payoff functions for roles a and b, and also infers/guesses his *transpersonal understanding* of the other's understanding.

Person i's direct understanding is purely based on his experiences. However, for his transpersonal understanding about j's understanding, we need a different kind of treatment reflecting postulates TP1 and TP2. Using those, we look for an experiential base for the other person's belief. These ideas are formulated in the following definition.

Definition 3.1 (Direct and Transpersonal Understandings). Let a memory kit $\kappa_i = \langle (s_a^o, s_b^o), (D_{ia}, D_{ib}), (h_{ia}, h_{ib}); (\rho_{ia}, \rho_{ib}) \rangle$ be given:

(1): The *direct understanding* (d-understanding) *of the situation from* κ_i by person i is given as $g^{ii}(\kappa_i) = (a, b, S_a^i, S_b^i, h_a^{ii}, h_b^{ii})$:

ID1^i: $S^i_r = \{s_r : (s_r; s_{-r}) \in D_{ia} \cup D_{ib} \text{ for some } s_{-r}\}$ for $r = a, b$;

ID2^{ii}: for $r = a, b$, h^{ii}_r is defined over $S^i_a \times S^i_b$ as follows:

$$h^{ii}_r(s_a, s_b) = \begin{cases} h_{ir}(s_a, s_b) & \text{if } (s_a, s_b) \in D_{ir} \\ \\ \theta_r & \text{otherwise,} \end{cases} \tag{3.1}$$

where θ_r is an exogenously given payoff value attached to every non-experienced (s_a, s_b).

(2): The *transpersonal understanding* (tp-understanding) *from* κ_i by person i for person j is given as $g^{ij}(\kappa_i) = (a, b, S^i_a, S^i_b, h^{ij}_a, h^{ij}_b)$, where only h^{ij}_a and h^{ij}_b are new and given as follows:

ID2^{ij}: for $r = a, b$, h^{ij}_r is defined over $S^i_a \times S^i_b$ by

$$h^{ij}_r(s_a, s_b) = \begin{cases} h_{ir}(s_a, s_b) & \text{if } (s_a, s_b) \in D_{ir} \text{ and } (s_a, s_b) \in D_{i(-r)} \\ \\ \theta_r & \text{otherwise.} \end{cases} \tag{3.2}$$

These understandings are deterministic: All the components of $g^{ii}(\kappa_i)$ and $g^{ij}(\kappa_i)$, except θ_r for the unexperienced part of $S^i_a \times S^i_b$, are determined from the components of κ_i. This differs from in Kaneko-Kline [16], [17], and [18]. This determinism comes from our restriction on the 2-role game with assumptions Ob1 and Ob2.

The definition of $g^{ii}(\kappa_i)$ is straightforward. He constructs his d-understanding as a 2-role game, based on his experiences. The symbol θ_r expresses an unknown (unexperienced) payoff, which is also assumed to be a real number and uniform over the experienced part. In ID1^i, the experienced actions are only taken into account. In ID2^{ii}, he constructs his observed payoff function. An example will be given presently. He notices more available actions in $S_r - S^i_r$, but he has no experiential information about the resulting outcomes from those actions. We assume that they are ignored in g^{ii} and also in g^{ij}.

The definition of $g^{ij}(\kappa_i)$ is less straightforward by its nature. Person i tries to analyze the experiences summarized in κ_i so as to obtain some information about the other's payoffs. By TP1, he projects his own experienced payoffs onto the other's thoughts. By TP2, however, he should only make this projection if he has reason to believe that the other has observed his payoff. In the top of (3.2), this projection is done for an experience (s_a, s_b) if and only if he experienced (s_a, s_b) from both roles.

Let us see (3.2) from the negative point of view: If at least one of $(s_a, s_b) \in D_{ir}$ and $(s_a, s_b) \in D_{i(-r)}$ does not hold, he cannot put the payoff value $h_{ir}(s_a, s_b)$ as $h^{ij}_r(s_a, s_b)$. Firstly, if i does not have the experience of (s_a, s_b) at role r , then the payoff information $h_{ir}(s_a, s_b)$ is not available to i, and *a fortiori*, he cannot project it onto j. Second, if $(s_a, s_b) \notin D_{i(-r)}$, then person i does not have reason to believe that j ever experienced

payoff $h_r(s_a, s_b)$, and he does not project his payoff experience, even if he has it, onto person j. Conversely, if both $(s_a, s_b) \in D_{ir}$ and $(s_a, s_b) \in D_{i(-r)}$ hold, he can project his experienced payoff onto the other person's thoughts.

The above requirement of having reason to believe is close to Lewis's [21] idea of person i having reason to believe that person j has also reason to believe the same. If we formulate the above argument as an epistemic logic system (cf., Kaneko [14]), we would examine this similarity more, which will be discussed in a separate paper. The argument here is entirely experiential, and in this sense, it is regarded as following the tradition from Mead [23].

Let us exemplify the above definitions with the examples from Section 2.2 assuming the regular actions $(s_a^o, s_b^o) = (\mathbf{s}_{a1}, \mathbf{s}_{b1})$:

(1)(Non-reciprocal Active Domain): Let (D_{1a}^N, D_{1b}^N) be given as the non-reciprocal

Table 3.1; g^{11}

	$\mathbf{s}_{b1}$
$\mathbf{s}_{a1}$	$(3, \theta_b)$
$\mathbf{s}_{a2}$	$(2, \theta_b)$
$\mathbf{s}_{a3}$	$(1, \theta_b)$

Table 3.2; g^{12}

	$\mathbf{s}_{b1}$
$\mathbf{s}_{a1}$	(θ_a, θ_b)
$\mathbf{s}_{a2}$	(θ_a, θ_b)
$\mathbf{s}_{a3}$	(θ_a, θ_b)

domain of (2.9), where we are considering only $G(1, 2)$. In this example, person 1's d-understanding $g^{11} = g^{11}(\kappa_1)$ is given as: $S_a^1 = \{\mathbf{s}_{a1}, \mathbf{s}_{a2}, \mathbf{s}_{a3}\}$ and $S_b^1 = \{\mathbf{s}_{b1}\}$ by ID1[1]. Since person 1 has experiences for role a, the payoffs $(h_a^{11}(s_a, s_b), h_b^{11}(s_a, s_b))$ become those described in Table 3.1. Since person 1 has no experiences with role b, his understanding of those payoffs $h_b^{11}(s_a, s_b)$ is simply θ_b.

Now, consider $g^{12}(\kappa_1)$. Person 1 has experienced the three pairs in D_{1a}^N, and from each pair, he guesses/infers that person 2 observes also these three pairs. Hence, person 1 can assume the same S_a^1 and S_b^1 for person 2, which corresponds to ID1[1]. But, now, person 1 has a real difficulty in guessing/inferring what person 2 could receive as payoffs from roles a and b. The easier part is $h_b^{12}(s_a, s_b) = \theta_b$ for role b since person 1 has no experiences with role b. The other equation $h_a^{12}(s_a, s_b) = \theta_a$ comes from $(s_a, s_b) \notin D_{1b}^N$: He infers from $(s_a, s_b) \notin D_{1b}^N$ that person 2 always plays role b and has no experiences with role a. Thus, person 1 should not project his experienced payoff onto 2's. In sum, $g^{12}(\kappa_1)$ is given as Table 3.2: Person 1 has no idea about person 2's understanding of payoffs.

(2)(Reciprocal Active Domain): Let (D_{1a}^A, D_{1b}^A) be given by the active domain of (2.10). By ID1[1], we have $S_a^1 = \{\mathbf{s}_{a1}, \mathbf{s}_{a2}, \mathbf{s}_{a3}\}$ and $S_b^1 = \{\mathbf{s}_{b1}, \mathbf{s}_{b2}, \mathbf{s}_{b3}\}$. Then, it follows from ID2[11] that (h_a^{11}, h_b^{11}) is given as Table 3.3. When person 1 is at b, he cannot guess/infer his own payoffs from trials of person 2 at a. Thus, he puts θ_b to the payoffs from trials in the first column of Table 3.3. For the same reason, he puts θ_a in Table

3.3 along the top row. The remaining four strategy combinations (s_a, s_b) belong neither D^A_{1a} nor D^A_{1b}, so he puts (θ_a, θ_b) in each case.

Table 3.3; g^{11}

$a\backslash b$	$\mathbf{s}_{b1}$	$\mathbf{s}_{b2}$	$\mathbf{s}_{b3}$
$\mathbf{s}_{a1}$	$(3,3)$	$(\theta_a, 2)$	$(\theta_a, 1)$
$\mathbf{s}_{a2}$	$(2, \theta_b)$	(θ_a, θ_b)	(θ_a, θ_b)
$\mathbf{s}_{a3}$	$(1, \theta_b)$	(θ_a, θ_b)	(θ_a, θ_b)

Table 3.4; g^{12}

$a\backslash b$	$\mathbf{s}_{b1}$	$\mathbf{s}_{b2}$	$\mathbf{s}_{b3}$
$\mathbf{s}_{a1}$	$(3,3)$	(θ_a, θ_b)	(θ_a, θ_b)
$\mathbf{s}_{a2}$	(θ_a, θ_b)	(θ_a, θ_b)	(θ_a, θ_b)
$\mathbf{s}_{a3}$	(θ_a, θ_b)	(θ_a, θ_b)	(θ_a, θ_b)

Person 1's tp-understanding g^{12} is even more restrictive as shown in Table 3.4. Let us see only how it comes that "$h^{12}_a(\mathbf{s}_{a2}, \mathbf{s}_{b1}) = \theta_a$". According to (D^A_{1a}, D^A_{1b}), person 1 has experienced the payoff $h_a(\mathbf{s}_{a2}, \mathbf{s}_{b1}) = 2$ and thus at least it would be possible for him to project this payoff onto person 2's. But since $(\mathbf{s}_{a2}, \mathbf{s}_{b1}) \notin D^A_{1b}$, he infers/guesses that 2 does not experience $(\mathbf{s}_{a2}, \mathbf{s}_{b1})$ at role a. So he puts $h^{12}_a(\mathbf{s}_{a2}, \mathbf{s}_{b1}) = \theta_a$.

The above observations hold more generally. Let $g^{ii}(\kappa_i) = (a, b, S^i_a, S^i_b, h^{ii}_a, h^{ii}_b)$ and $g^{ij}(\kappa_i) = (a, b, S^i_a, S^i_b, h^{ij}_a, h^{ij}_b)$ be the d- and tp-understandings.

Lemma 3.1: Let $\rho_{ir} = 1$. Then, $h^{ii}_{-r}(s_a, s_b) = h^{ij}_{-r}(s_a, s_b) = \theta_{-r}$ and $h^{ij}_r(s_a, s_b) = \theta_r$ for all $(s_a, s_b) \in S^i_a \times S^i_b$.

Proof. Since $\rho_{ir} = 1$, we have $D_{i(-r)} = \emptyset$ by (2.6). By (3.1) and (3.2), we have the stated equations. ■

When the situation is reciprocal and when person 1 is equally sensitive to the experiences caused by person 2, he has the active-passive domain.

(3):(Reciprocal Active-Passive Domain): Let $D^{AP}_1 = (D^{AP}_{1a}, D^{AP}_{1b})$ be the domains described by (2.11). By ID^1, we have $S^1_a = \{\mathbf{s}_{a1}, \mathbf{s}_{a2}, \mathbf{s}_{a3}\}$ and $S^1_b = \{\mathbf{s}_{b1}, \mathbf{s}_{b2}, \mathbf{s}_{b3}\}$. But

Table 3.5; g^{11} and g^{12}

$a\backslash b$	$\mathbf{s}_{b1}$	$\mathbf{s}_{b2}$	$\mathbf{s}_{b3}$
$\mathbf{s}_{a1}$	$(3,3)$	$(10,2)$	$(3,1)$
$\mathbf{s}_{a2}$	$(2,10)$	(θ_a, θ_b)	(θ_a, θ_b)
$\mathbf{s}_{a3}$	$(1,3)$	(θ_a, θ_b)	(θ_a, θ_b)

the payoff functions (h^{11}_a, h^{11}_b) given in Table 3.5 are different from those in Table 3.3. Indeed, $h^{11}_b(\mathbf{s}_{a2}, \mathbf{s}_{b1}) = 10$ by ID2^{11}, since $(\mathbf{s}_{a2}, \mathbf{s}_{b1}) \in D^{AP}_{1b}$. Person 1 is also sensitive with passive experiences from person 2's active deviations. This means that $D_{1a} = D_{1b}$. In this example, the payoff functions (h^{12}_a, h^{12}_b) are the same as Table 3.5. Person 1 has had each experience along the top row and down the first column from the perspective of each role. Thus, he can and does project his experiences onto the other person. Only the joint trials are excluded as they are outside his domains of accumulation.

This internal reciprocity and coincidence will be important in our later analysis. We will give one theorem on this, which states that internal reciprocity (2.8) is necessary and sufficient for coincidence of a person's direct and transpersonal understandings up to the active and passive experiences. Let $g^{ii}(\kappa_i)$ and $g^{ij}(\kappa_i)$ be the d- and tp-understandings from a memory kit κ_i.

Theorem 3.2 (1): If (D_{ia}, D_{ib}) is internally reciprocal, i.e., $\text{Proj}(D_{ia}) = \text{Proj}(D_{ib})$, then $h_r^{ii}(s_a, s_b) = h_r^{ij}(s_a, s_b) = h_r(s_a, s_b)$ for all $(s_a, s_b) \in \text{Proj}(S_a^i \times S_b^i)$.

(2)(Internal Coincidence): $g^{ii}(\kappa_i)$ coincides with $g^{ij}(\kappa_i)$ up to the active/passive experiences, i.e., $h_r^{ii}(s_a, s_b) = h_r^{ij}(s_a, s_b)$ for all $(s_a, s_b) \in \text{Proj}(S_a^i \times S_b^i)$ and all θ_a, θ_b if and only if (D_{ia}, D_{ib}) is internally reciprocal.

Proof. (1): Suppose $\text{Proj}(D_{ia}) = \text{Proj}(D_{ib})$. Then, $\text{Proj}(S_a^i \times S_b^i) = \text{Proj}(D_{ia}) = \text{Proj}(D_{ib})$ by (2.3). Hence, the stated equations follow (3.1) and (3.2).

(2): The if part is already proved in (1). Consider the only-if part. It suffices to show that $(s_a, s_b) \in \text{Proj}(D_{ir})$ implies $(s_a, s_b) \in D_{i(-r)}$. Let $(s_a, s_b) \in \text{Proj}(D_{ir})$. Then, $(s_a, s_b) \in \text{Proj}(S_a^i \times S_b^i)$, which means $h_r^{ij}(s_a, s_b) = h_r^{ii}(s_a, s_b)$. Then, since $(s_a, s_b) \in \text{Proj}(D_{ir})$, we have $h_r^{ii}(s_a, s_b) = h_r(s_a, s_b)$ by (3.1). If $(s_a, s_b) \notin D_{i(-r)}$, then $h_r^{ij}(s_a, s_b) = \theta_r$ by (3.2), and for some choice of θ_r, we have $h_r^{ii}(s_a, s_b) \neq h_r^{ij}(s_a, s_b)$, a contradiction. Thus, $(s_a, s_b) \in D_{i(-r)}$. ■

4. Inductively Derived Views and Their Use for Behavioral Revision

4.1. Inductively Derived View

The understandings $g^{ii}(\kappa_i)$ and $g^{ij}(\kappa_i)$ do not take the regular actions (s_a^o, s_b^o) and the frequency weights (ρ_{ia}, ρ_{ib}) into account. The inductively derived view is defined by adding these two components.

Since each person acts roles a or b at different times and with different frequencies, we need weighted payoff functions. Since the weighted payoff functions in person i's mind depend on the actions by each person at each role, we introduce the expression $[s_a, s_b]_r$ to mean that person i takes role r in playing (s_a, s_b). The importance of this new expression will become clear when we consider deviations in Sections 4.2 and 5.

Definition 4.1. The *inductively derived view (i.d.view) from the memory kit* $\kappa_i = \langle (s_a^o, s_b^o), (D_{ia}, D_{ib}), (h_{ia}, h_{ib}); (\rho_{ia}, \rho_{ib}) \rangle$ is given as $\Gamma^i = \langle (s_a^o, s_b^o), (S_a^i, S_b^i), (\rho_{ia}, \rho_{ib}), H^{ii}, H^{ij} \rangle$, where the additional H^{ii} and H^{ij} are the weighted payoff functions given as follows: for all $([s_a, s_b], [t_a, t_b]) \in (S_a^i \times S_b^i)^2$,

$$H^{ii}([s_a, s_b]_a, [t_a, t_b]_b) = \rho_{ia} h_a^{ii}(s_a, s_b) + \rho_{ib} h_b^{ii}(t_a, t_b); \tag{4.1}$$

$$H^{ij}([s_a, s_b]_a, [t_a, t_b]_b) = \rho_{ia} h_b^{ij}(s_a, s_b) + \rho_{ib} h_a^{ij}(t_a, t_b). \tag{4.2}$$

The payoff functions H^{ii} and H^{ij} are considered for persons i and j in the mind of person i. The payoffs are taken as weighted averages of the payoffs of g^{ii} and g^{ij} with the frequency weights (ρ_{ia}, ρ_{ib}). We should notice a break in symmetry in (4.1) and (4.2): In (4.2), when person i plays role a, person j plays role b; hence, the first term of the right-hand side of (4.2) means that person j takes role b with frequency ρ_{ia}. The second term has the parallel meaning.

The definition of the i.d.view Γ^i has various differences from those given in Kaneko-Kline [16], [17] and [18]. One apparent difference is that the definition is given to a strategic game but not an extensive game (or an information protocol). This also makes the view here deterministic as $g^{ii}(\kappa_i)$ and $g^{ij}(\kappa_i)$. But it is the most important point to include the weighted payoffs coming from role-switching.

The sums with frequency weights are based on the frequentist interpretation of expected utility theory, which is close to the original interpretation by von Neumann-Morgenstern [28]. See Hu [12] for a direct approach to expected utility theory from the frequentist perspective.

As noted in Section 2, the frequency weights should be interpreted as rough descriptions of past occurrences of roles in the mind of person i. By postulates BH1, BH2, EP1 and EP2 in Section 2.3, when the objective frequency of taking role b is small close to 0, person i effectively takes it to be 0, i.e., $\rho_{ib} = 0$. By (2.6), his domain D_{ib} of accumulated experiences is empty, and the situation is effectively non-reciprocal.

Following this interpretation, ρ_{ia} does not vary continuously, but take rather discrete values, as the memory kit κ_i (in particular, the domains of accumulated experiences D_{ia} and D_{ib}) is finitistic and discrete. At present, we do not a study of relationships between the objective frequency of role-switching and the subjective evaluation of ρ_{ia} (and D_{ia} and D_{ib}). To study such relationships, computer situations such as the one in Akiyama et al. [1] will play a crucial role, since they must be of truly finite nature.

In our interpretation, the frequency weight ρ_{ia} may take only finite and discrete values; candidates are

$$\alpha_{ia0} = 0 < \alpha_{ia1} < \ ... \ < \alpha_{iam} = 1 \tag{4.3}$$

In the following, we assume that this list includes 1/2. When $\rho_{ia} = \alpha_{iam} = 1$, person i takes role a exclusively, at least in his mind. In this case, $D_{ib} = \emptyset$ by (2.6), $H^{ii}([s_a, s_b]_a, [t_a, t_b]_b)$ is reduced into $h^{ii}_a(s_a, s_b)$, and $H^{ij}([s_a, s_b]_a, [t_a, t_b]_b)$ becomes the constant function taking the value θ_b. In this sense, we may regard the i.d.view Γ^i for $\rho_{ia} = 1$ as not including the other's thoughts.

4.2. Partial vs. Full Use of the I.D.View

Now, consider how person i uses the i.d.view $\Gamma^i = \langle (s^o_a, s^o_b), (S^i_a, S^i_b), (\rho_{ia}, \rho_{ib}), H^{ii}, H^{ij} \rangle$. It includes the tp-understanding of the other person's payoffs in addition to his own d-understanding. When person i uses Γ^i for his decision making, he would face the

problem of whether or not he should use the tp-understanding. We have the following two cases:

C0(Partial Use): Person i uses only the payoff function H^{ii} (assuming some reactions of person j).

C1(Full Use): Person i uses not only the payoff function H^{ii} but also H^{ij} to in order predict how person j will act (or react).

Either C0 or C1 may be taken as a decision criterion for person i. The choice of C0 or C1 is logically independent of the degree of reciprocity, though it likely to be some correlation between them. For example, when $\rho_{ir} = 0$ or 1, person i cannot choose C1 effectively. As the degree of reciprocity increases, the criterion C1 may emerge in a player's mind. Here, we consider first C0 and then we go to the full use of Γ^i.

In C0, person i can maximize his weighted payoff H^{ii} by choosing his action from the assigned role in one play of the game. Since he uses only H^{ii}, he needs some assumption about the other person's action or reaction to his change. An assumption for C0 is:

$(*)$: person j sticks to the regular action.

In this case, person i may choose a maximum point in S_r^i against the regular action s^o_{-r}. If the present regular actions are free from such behavior revisions, then the regular action pair (s_a^o, s_b^o) must be a Nash equilibrium in the d-understanding g^{ii}. We do not pursue what happen in this case; the main aim of the present paper is to study the reciprocal case and the full use of Γ^i.

Consider the case where person i uses H^{ij}. We formulate one concept of an equilibrium expressing the idea that person i thinks about the present regular behavior as a satisfactory result even taking into account the other's thinking in his mind. It is a salient point different from the above assumption $(*)$ that coordination may be considered in the mind of person i. Since this involves a new and subtle argument about comparisons between the regular action s_r^o and another action s_r in S_r^i, we start with a clear-cut case. It should be kept in mind that the entire argument is made inside the mind of person i. In the following, we let $r = a$.

Now, we consider a possible deviation by person i using H^{ii} and H^{ij} : Person i evaluates his action in terms of H^{ii} relative to his regular action, and predicts what person j would think, by his H^{ij}. Now, suppose that

$$H^{ii}([s_a^o, s_b^o]_a, [s_a^o, s_b^o]_b) < H^{ii}([s_a, s_b^o]_a, [s_a, s_b^o]_b). \tag{4.4}$$

This means that i would get a higher weighted payoff by deviating from s_a^o to s_a and assuming that person j also deviates from s_a^o to s_a. This assumption part is expressed by $[s_a, s_b^o]_b$ meaning that person j taking role a chooses action s_a. This is the main difference from $(*)$. An apparent question is why person i can make this assumption that person j will choose action s_a.

The answer is as follows: We require the parallel inequality:

$$H^{ij}([s_a^o, s_b^o]_a, [s_a^o, s_b^o]_b) < H^{ij}([s_a, s_b^o]_a, [s_a, s_b^o]_b). \tag{4.5}$$

If this holds, person i thinks that person j thinks in the same manner as (4.4) and evaluates s_a as better than s_a^o. Person i now can believe that the deviation s_a from the regular action s_a^o gives a higher payoff for both persons, and that person j thinks in the same manner.

In (4.4) and (4.5), we considered only a unilateral derivation s_a from s_a^o, and we can also consider another parallel unilateral derivation s_b from s_b^o. Mathematically, we may consider even a joint deviation (s_a, s_b) from (s_a^o, s_b^o) satisfying (4.4) and (4.5). However, this requires some direct coordination or communication between the persons. In our context, we have a lot of possible ways of coordination or communication. It would be better to separate studies of these possibilities from the present research, which should be discussed in a separate paper.

Let us return to the unilateral deviation in (4.4) and (4.5). This deviation needs only one person to deviate first. That is, the deviation process can be expressed as

$$\rightarrow \quad \begin{pmatrix} 1, & 2 \\ s_a^o, & s_b^o \end{pmatrix} \rightarrow \begin{pmatrix} 1, & 2 \\ s_a, & s_b^o \end{pmatrix} \rightarrow \begin{pmatrix} 2, & 1 \\ s_a, & s_b^o \end{pmatrix} \rightarrow \begin{pmatrix} 2, & 1 \\ s_a, & s_b^o \end{pmatrix} \rightarrow$$

Fig.4.1

That is, suppose (s_a^o, s_b^o) is the current regular pair. Next, suppose that 1 deviates from s_a^o to a mutually beneficial s_a, which is the second left state in Fig.4.1. Then person 2 will observe this deviation, and when 2 is assigned role a, he follows 1's mutually beneficial deviation to s_a, which is describe as the third state in Fig.4.1.

We now follow the standard idea of an equilibrium to be free from such a deviation. We formulate it as follows:

Definition 4.2 (Weak I.C.Equilibrium). We say that the regular pair (s_a^o, s_b^o) is *a weak intrapersonal coordination equilibrium* (*weak i.c.equilibrium*) *in* Γ^i iff there is neither $s_a \in S_a^i$ satisfying (4.4) and (4.5) nor $s_b \in S_b^i$ satisfying (4.4) and (4.5) with the replacements of s_a by s_b.

We present the following existence theorem. To avoid a messy presentation, we strengthen (2.6): for $r = a, b$,

$$\rho_{ir} = 0 \text{ if and only if } D_{ir} = \emptyset. \tag{4.6}$$

That is, this requires the converse of (2.6).

Theorem 4.1 (Existence of a Weak I.C.Equilibrium). We assume that

$$\theta_r \leq \min_{(s_a, s_b) \in S_a \times S_b} h_r(s_a, s_b) \text{ for } r = a, b. \tag{4.7}$$

For any frequency weights (ρ_{ia}, ρ_{ib}), there is a pair (s_a^*, s_b^*) in $S_a \times S_b$ such that for any memory kit $\kappa_i = \langle (s_a^o, s_b^o), (D_{ia}, D_{ib}), (h_{ia}, h_{ib}); (\rho_{ia}, \rho_{ib}) \rangle$ satisfying (4.6) with $(s_a^o, s_b^o) = (s_a^*, s_b^*)$, (s_a^o, s_b^o) is a weak i.c.equilibrium in $\Gamma^i(\kappa_i)$.

Proof. Let (s_a^*, s_b^*) be a maximizer of $\rho_{ia} h_a(s_a, s_b) + \rho_{ib} h_b(s_a, s_b)$ over $S_a \times S_b$. There are three cases to be considered: (a) $(s_a^*, s_b^*) \in D_{ia} \cap D_{ib}$; (b) $(s_a^*, s_b^*) \in D_{ia} - D_{ib}$; and (c) $(s_a^*, s_b^*) \in D_{ib} - D_{ia}$. Cases (b) and (c) are symmetric. We consider (a) and (b).

Consider (a). Then, $h_r^{ii}(s_a^*, s_b^*) = h_r(s_a^*, s_b^*)$ for $r = a, b$, which implies $H^{ii}([s_a^*, s_b^*]_a, [s_a^*, s_b^*]_b)$ $= \rho_{ia} h_a(s_a^*, s_b^*) + \rho_{ib} h_b(s_a^*, s_b^*)$, which is a maximum of $\rho_{ia} h_a(s_a, s_b) + \rho_{ib} h_b(s_a, s_b)$ over $S_a \times S_b$. Now, let $s_r \in S_r^i$. Then, by (4.7), we have $H^{ii}([s_a^*, s_b^*]_a, [s_a^*, s_b^*]_b) \geq H^{ii}([s_r; s_{-r}^*]_a,$ $[s_r, s_b^*]_b)$.

Consider case (b). In this case, $(s_a^o, s_b^o) = (s_a^*, s_b^*) \notin D_{ib}$. By Lemma 2.1, $D_{ib} = \emptyset$. By (4.6), we have $\rho_{ib} = 0$. Thus, (s_a^*, s_b^*) is a maximizer of $h_a(s_a, s_b)$ over $S_a \times S_b$. Then, based on (4.7), we have $H^{ii}([s_a^*, s_b^*]_a, [s_a^*, s_b^*]_b) \geq H^{ii}([s_r; s_{-r}^*]_a, [s_r, s_b^*]_b)$ for any s_r. ■

When each role has only a few actions, all actions could be experienced from each role. Condition (4.7) may be irrelevant for the existence result.

This definition has some difficulties. One is that it is conceptually poor in the non-reciprocal case, i.e., $\rho_{ir} = 0$ or 1, which may be found by recalling Lemma 3.1 and will be shown more explicitly in Section 5.2. A possible remedy will be discussed there.

Another difficulty is about the deviation story given in (4.4) and (4.5). It appears to define an equilibrium from an external viewpoint, i.e., a deviation of (4.4) and (4.5) is interpreted as a deviation from the viewpoint of an outsider. We would like to concentrate on the internal thinking of person i.

Yet, another difficulty, related to the previous one, is that a weak i.c.equilibrium does not require utility maximization of each person. It may even happen that a regular behavior (s_a^o, s_b^o) is a weak i.c.equilibrium in i's mind simply because he worries that any profitable deviation to himself in the sense of (4.4) will harm person j. In the next section we will consider a strengthening of the weak i.c.equilibrium by requiring utility maximization in the mind of each person.

5. Intrapersonal Coordination Equilibrium

5.1. Intrapersonal Coordination Equilibria through the I.D.View Γ^i

Here, we give a strengthening of an weak i.c.equilibrium. Let $\Gamma^i = \langle (s_a^o, s_b^o), (S_a^i, S_b^i), (\rho_{ia}, \rho_{ib}),$ $H^{ii}, H^{ij} \rangle$ be the i.d.view derived from the memory kit κ_i.

Definition 5.1 (I.C.Equilibrium). We say that the regular pair (s_a^o, s_b^o) is a *an intrapersonal coordination equilibrium* (i.c.equilibrium) *in* Γ^i iff for all $s_a \in S_a^i$,

$$\begin{aligned} H^{ii}([s_a^o, s_b^o]_a, [s_a^o, s_b^o]_b) &\geq H^{ii}([s_a, s_b^o]_a, [s_a, s_b^o]_b) \\ H^{ij}([s_a^o, s_b^o]_a, [s_a^o, s_b^o]_b) &\geq H^{ij}([s_a, s_b^o]_a, [s_a, s_b^o]_b); \end{aligned} \tag{5.1}$$

and for all $s_b \in S_b^i$,

$$\begin{aligned} H^{ii}([s_a^o, s_b^o]_a, [s_a^o, s_b^o]_b) &\geq H^{ii}([s_a^o, s_b]_a, [s_a^o, s_b]_b) \\ H^{ij}([s_a^o, s_b^o]_a, [s_a^o, s_b^o]_b) &\geq H^{ij}([s_a^o, s_b]_a, [s_a^o, s_b]_b). \end{aligned} \tag{5.2}$$

That is, person i thinks, based on his i.d.view Γ^i, that s_a^o gives higher payoff to both persons 1 and 2 than any other action $s_a \in S_a^i$ with their coordination, and s_b^o has the same property. As emphasized in Section 4.2, this coordination is considered in the mind of person i.

Inequalities (5.1) and/or (5.2) may include the case where we have the strict inequality for H^{ii} but the equality for H^{ij}; this may hold if person i's tp-understanding is trivial or poor, e.g., $\rho_{ir} = 0$ or 1. Thus, although the definition of an i.c.equilibrium is given by two inequality systems for each role, it includes cases where he has a poor tp-understanding or even his d-understanding is poor, e.g., S_a^i and S_b^i consist only of regular actions.

First, we show that an i.c.equilibrium is a strengthening of a weak i.c.equilibrium, and that they coincide in the fully reciprocal case.

Lemma 5.1.(1): If (s_a^o, s_b^o) is an i.c.equilibrium, then it is a weak i.c.equilibrium.

(2): Let (D_{ia}, D_{ib}) be internally reciprocal and $(\rho_{ia}, \rho_{ib}) = (\frac{1}{2}, \frac{1}{2})$. Then, if (s_a^o, s_b^o) is a weak i.c.equilibrium, then it is an i.c.equilibrium.

Proof. Assertion (1) is straightforward from (5.1) and (5.2).

Consider (2). Let (s_a^o, s_b^o) be a weak i.c.equilibrium. Then, since (D_{ia}, D_{ib}) is internally reciprocal, it follows from Theorem 3.1 and (2.3) that for all $(s_a, s_b) \in S_a^i \times S_b^i$,

$$\text{if } s_a = s_a^o \text{ or } s_b = s_b^o, \text{ then } h_r^{ii}(s_a, s_b) = h_r^{ij}(s_a, s_b) = h_r(s_a, s_b).$$

Since (s_a^o, s_b^o) is a weak i.c.equilibrium, we have, for each s_a, at least (4.4) and (4.5) does not hold. In either case, we have

$$\frac{1}{2}h_a(s_a^o, s_b^o) + \frac{1}{2}h_b(s_a^o, s_b^o) \geq \frac{1}{2}h_a(s_a, s_b^o) + \frac{1}{2}h_b(s_a, s_b^o).$$

This is (5.1) for internally reciprocal (D_{ia}, D_{ib}) and $(\rho_{ia}, \rho_{ib}) = (\frac{1}{2}, \frac{1}{2})$. Similarly, we have (5.2). ■

As shown in Theorem 4.1, a weak i.c.equilibrium exists but it behaves poorly in the non-reciprocal case. Perhaps, we should keep both equilibrium concepts in our mind, but we will focus on an i.c.equilibrium in the sense of Definition 5.1 in the remainder of the paper. One reason is that it maintains the utility maximization, as mentioned at the end of Section 4.2.

Before going to the next section, we will give one more definition. Suppose that person $i = 1, 2$ has an i.d.view $\Gamma^i = \langle (s_a^o, s_b^o), (S_a^i, S_b^i), (\rho_{ia}, \rho_{ib}), H^{ii}, H^{ij} \rangle$ derived from a

memory kit $\kappa_i = \langle (s_a^o, s_b^o), (D_{ia}, D_{ib}), (h_{ia}, h_{ib}); (\rho_{ia}, \rho_{ib}) \rangle$.

Definition 5.2 (Mutual I.C.Equilibrium). We say that the pair (s_a^o, s_b^o) of regular actions is a *mutual i.c.equilibrium* iff it is an i.c.equilibrium for both Γ^1 and Γ^2.

Our goal is to study the 2-person game situation and the interactions of the persons there, rather than just to consider an i.c.equilibrium from the viewpoint of one person. Therefore, our final objective is to study a mutual i.c.equilibrium. Nevertheless, since it is required to be an i.c.equilibrium for each person, a research method becomes to study first an i.c.equilibrium. Then, we will synthesize it to a mutual i.c.equilibrium.

5.2. Non-reciprocal Domains and Reciprocal Active Domain

There is a spectrum of reciprocal degrees of switching roles between the two persons. The non-reciprocal domains (2.9) and active domains (2.10) are located at the lowest side of this spectrum, while the fully reciprocal domains are located at the other extreme. It is our intention to show that cooperation is emerging as the reciprocal degree is increasing. To show this, we first show that at the lowest end, no cooperation occurs, more concretely, for the non-reciprocal domains and active domains, the i.c.equilibrium yields non-cooperative outcomes. In Section 6, we will consider the other extreme case of the spectrum of reciprocal degrees.

The following theorem is a simple observation of an i.c.equilibrium and a weak i.c.equilibrium in the non-reciprocal case.

Theorem 5.2 (Non-reciprocal Case): **(1)**: Let $\rho_{ir} = 1$. Then, the pair (s_a^o, s_b^o) of regular actions is an i.c.equilibrium in an i.d.view Γ^i if and only if it is a Nash equilibrium in person i's d-understanding g^{ii}.

(2): For any i.d.view $\Gamma^i = \langle (s_a^o, s_b^o), (S_a^i, S_b^i), (\rho_{ia}, \rho_{ib}), H^{ii}, H^{ij} \rangle$ with $\rho_{ir} = 1$, (s_a^o, s_b^o) is a with $\rho_{ir} = 1$, (s_a^o, s_b^o) is a weak i.c.equilibrium.

Proof. (1): Let s_r be an arbitrary element in S_r^i. Let (s_a^o, s_b^o) be an i.c.equilibrium in Γ^i. Then, using

$$\begin{aligned} \rho_{ir} h_r^{ii}(s_r^o; s_{-r}^o) + (1-\rho_{ir}) h_{-r}^{ii}(s_r^o; s_{-r}^o) &= H^{ii}([s_a^o, s_b^o]_a, [s_a^o, s_b^o]_b) \geq \\ H^{ii}([s_r; s_{-r}^o]_a, [s_r; s_{-r}^o]_b) &= \rho_{ir} h_r^{ii}(s_r; s_{-r}^o) + (1-\rho_{ir}) h_{-r}^{ii}(s_r; s_{-r}^o). \end{aligned} \tag{5.3}$$

Now, since $D_{i(-r)} = \emptyset$, we have $(s_r; s_{-r}^o) \in D_{ir}$ for all $s_r \in S_r^i$. Hence $h_r^{ii}(s_r; s_{-r}^o) = h_r(s_r; s_{-r}^o)$ for all $s_r \in S_r^i$. In particular, $(s_r^o; s_{-r}^o) \in D_{ir}$ by Lemma 2.1. By these and (5.3), we have $h_r(s_r^o; s_{-r}^o) \geq h_r(s_r; s_{-r}^o)$ using $\rho_{ir} = 1$. By (1), (s_a^o, s_b^o) is a Nash equilibrium in g^{ii}. Tracing the argument back, we have the only-if part, i.e., if (s_a^o, s_b^o) is a Nash equilibrium in g^{ii}, then (s_a^o, s_b^o) be an i.c.equilibrium in Γ^i.

(2): Lemma 3.1 states that h_r^{ij} is constant over $S_a^i \times S_b^i$. Hence, (4.5) does not hold, and we have the assertion. ■

In the above theorem, a Nash equilibrium in person i's d-understanding g^{ii} is simply a payoff maximization point in the base game G with the fixed s^o_{-r}. Hence, we have the following corollary.

Corollary 5.3 (Mutual I.C.Equilibrium in the Non-reciprocal Domains): Let $\rho_{ir} = \rho_{j(-r)} = 1$, $S^i_r = S_r$, $S^j_{-r} = S_{-r}$, and (s^o_a, s^o_b) the pair of regular actions Then, (s^o_a, s^o_b) is a mutual i.c.equilibrium if and only if it is a Nash equilibrium in the base game G.

Assertion (2) of Theorem 5.2 states that a weak i.c.equilibrium becomes poor when ρ_{ir} is 0 or 1. On the other hand, an i.c.equilibrium may disappear for some other range of ρ_{ir}, which will be explained in Section 6. To a great extent, these are complementary equilibrium concepts. It is the salient point, stated by Lemma 5.1, that they coincide when $(\rho_{ia}, \rho_{ib}) = (1/2, 1/2)$ and the domains (D_{ia}, D_{ib}) are internally reciprocal. In fact, we will show that in that case, the cooperation outcome results.

In fact, assertion (2) could be removed by choosing the slightly stronger form of a weak i.c.equilibrium. Since this fact may help the reader understand some other possible definition of our equilibrium concept, we give it as a remark.

Remark 5.1. A possible amendment of a weak i.c.equilibrium is: We replace (4.5) by a weak inequality, i.e., it is free from a deviation of (4.4) and the weaker form of (4.5). This change would eliminate the poorness in (2) of Theorem 5.2 and make (1) to hold; and simultaneously we can keep the existence theorem (Theorem 5.1). Why we adopt the present i.c.equilibrium as well as mention the weak i.c.equilibrium is that the amendment rely upon a subtle story of deviations for both persons.

Before going to Section 6, we just mention, without a proof, the behavior of an i.c.equilibrium in the reciprocal active domains.

Theorem 5.4 (Reciprocal Active Domain): Let (D^A_{ia}, D^A_{ib}) be the reciprocal active domain for person $i = 1, 2$ defined in (2.10) with the regular actions (s^o_a, s^o_b). Suppose that $\theta_a \leq h_a(s^o_a, s^o_b)$ and $\theta_b \leq h_b(s^o_a, s^o_b)$. Then the following two statements hold:

(1): If (s^o_a, s^o_b) is a Nash equilibrium in the 2-role strategic game $G = (a, b, S_a, S_b, h_a, h_b)$, then it is an i.c.equilibrium.

(2): Suppose that $h_r(s^o_a, s^o_b) = \theta_r$ for $r = a, b$. Then the converse of (1) holds.

6. Intrapersonal Coordination Equilibrium for Reciprocal Domains

The theorems in Section 5.2 stated that in the non-reciprocal case, the i.c.equilibrium results as a noncooperative outcome, while the weak i.c.equilibrium does not exclude any outcome as a candidate. In this section, we will show that both equilibrium concepts suggest the cooperative outcome when domains (D_{ia}, D_{ib}) are internally reciprocal and

$(\rho_{ia}, \rho_{ib}) = (1/2, 1/2)$. We interpret this as the emergence of cooperation from a sufficient degree of reciprocity. As already mentioned, the frequency weights should be interpreted as a rough description. First, we will give a result when $(\rho_{ia}, \rho_{ib}) = (1/2, 1/2)$, and then mention how this assumption is weakened with a rough interpretation of $(\rho_{ia}, \rho_{ib}) = (1/2, 1/2)$.

Let $\Gamma^i = \langle (s_a^o, s_b^o), (S_a^i, S_b^i), (\rho_{ia}, \rho_{ib}), H^{ii}, H^{ij} \rangle$ be an i.d.view.

Theorem 6.1.(Utilitarian Condition): Suppose that domains (D_{ia}, D_{ib}) are internally reciprocal and $(\rho_{ia}, \rho_{ib}) = (1/2, 1/2)$. Then, the pair (s_a^o, s_b^o) of regular actions is an i.c.equilibrium for Γ^i if and only if

$$h_a(s_a^o, s_b^o) + h_b(s_a^o, s_b^o) \geq h_a(s_a, s_b^o) + h_b(s_a, s_b^o) \text{ for all } s_a \in S_a^i; \tag{6.1}$$

$$h_a(s_a^o, s_b^o) + h_b(s_a^o, s_b^o) \geq h_a(s_a^o, s_b) + h_b(s_a^o, s_b) \text{ for all } s_a \in S_b^i. \tag{6.2}$$

Proof. By (6.3), there is an $\alpha' \in (0, \frac{1}{2})$ such that for any $\hat{\rho}_{ia} \in (\frac{1}{2} - \alpha, \frac{1}{2} + \alpha)$,

$$\hat{\rho}_{ia} h_a(s_a^o, s_b^o) + \hat{\rho}_{ib} h_b(s_a^o, s_b^o) \geq \hat{\rho}_{ia} h_a(s_a, s_b^o) + \hat{\rho}_{ib} h_b(s_a, s_b^o) \text{ for all } s_a \in S_a^i;$$

$$\hat{\rho}_{ia} h_b(s_a^o, s_b^o) + \hat{\rho}_{ia} h_a(s_a^o, s_b^o) \geq \hat{\rho}_{ia} h_b(s_a, s_b^o) + \hat{\rho}_{ib} h_a(s_a^o, s_b) \text{ for all } s_a \in S_a^i,$$

which correspond to (5.1). By (6.4), we can find some $\alpha'' \in (0, \frac{1}{2})$ so that the inequalities corresponding to (5.2) hold. Let $\alpha = \min(\alpha', \alpha'')$. Since (D_{ia}, D_{ib}) are internally reciprocal, we have $h_r^{ii}(s_a, s_b) = h_r^{ij}(s_a, s_b) = h_r(s_a, s_b)$ for all $(s_a, s_b) \in \text{Proj}(S_a^i \times S_b^i)$ by Theorem 3.2.(1). Hence, these inequalities imply (5.1) and (5.2). ■

By Lemma 5.1, this theorem holds also for a weak i.c.equilibrium (in fact, the equilibrium concept mentioned in Remark 5.1 coincides with them, too). Hence, both concepts suggest that cooperation results in the fully reciprocal case, while they behave quite differently in the non-reciprocal case. The title "Utilitarian Condition" of the theorem will be explained in Section 7.

The above theorem holds even for some interval of ρ_{ia} centered at 1/2 (and also for the weak i.c.equilibrium) under some additional condition. Again, $\Gamma^i = \langle (s_a^o, s_b^o), (S_a^i, S_b^i), (\rho_{ia}, \rho_{ib}), H^{ii}, H^{ij} \rangle$ is assumed to be an i.d.view.

Theorem 6.2.(Rough Weights): Suppose that domains (D_{ia}, D_{ib}) are internally reciprocal and that

$$h_a(s_a^o, s_b^o) + h_b(s_a^o, s_b^o) > h_a(s_a, s_b^o) + h_b(s_a, s_b^o) \text{ for all } s_a \in S_a^i \backslash \{s_a^o\}; \tag{6.3}$$

$$h_a(s_a^o, s_b^o) + h_b(s_a^o, s_b^o) > h_a(s_a^o, s_b) + h_b(s_a^o, s_b) \text{ for all } s_b \in S_b^i \backslash \{s_b^o\}. \tag{6.4}$$

Then there is a $\alpha \in (0, \frac{1}{2})$ such that for any $\hat{\rho}_{ia} \in (\frac{1}{2} - \alpha, \frac{1}{2} + \alpha)$, (s_a^o, s_b^o) is an i.c.equilibrium for $\hat{\Gamma}^i$ obtained from Γ^i by the replacement of (ρ_{ia}, ρ_{ib}) with $(\hat{\rho}_{ia}, \hat{\rho}_{ib}) =$

$(\hat{\rho}_{ia}, 1-\hat{\rho}_{ia})$.

Proof. It follows (6.3) that for some $\alpha \in (0, \frac{1}{2})$, for any $\hat{\rho}_{ia} \in (\frac{1}{2}-\alpha, \frac{1}{2}+\alpha)$,

$$\hat{\rho}_{ia}h_a(s_a^o, s_b^o) + \hat{\rho}_{ib}h_b(s_a^o, s_b^o) \geq \hat{\rho}_{ia}h_a(s_a, s_b^o) + \hat{\rho}_{ib}h_b(s_a, s_b^o) \text{ for all } s_a \in S_a^i;$$

$$\hat{\rho}_{ia}h_b(s_a^o, s_b^o) + \hat{\rho}_{ia}h_a(s_a^o, s_b^o) \geq \hat{\rho}_{ia}h_b(s_a, s_b^o) + \hat{\rho}_{ib}h_a(s_a^o, s_b) \text{ for all } s_a \in S_a^i,$$

which correspond to (5.1), and we can have, (6.4), the inequalities corresponding to (5.2). Since (D_{ia}, D_{ib}) are internally reciprocal, we have $h_r^{ii}(s_a, s_b) = h_r^{ij}(s_a, s_b) = h_r(s_a, s_b)$ for all $(s_a, s_b) \in \text{Proj}(S_a^i \times S_b^i)$ by Theorem 3.2.(1). Hence, these inequalities imply (5.1) and (5.2). ■

Now, we have seen that we do not need to take $(\rho_{ia}, \rho_{ib}) = (\frac{1}{2}, \frac{1}{2})$ as a very accurate requirement. In the following, we forget rough weights, and state the existence theorem of an i.c.equilibrium for internally reciprocal domains and $(\rho_{ia}, \rho_{ib}) = (\frac{1}{2}, \frac{1}{2})$.

Theorem 6.3 (Existence of an I.C.Equilibrium): Let $(\rho_{ia}, \rho_{ib}) = (\frac{1}{2}, \frac{1}{2})$. Then, there is a pair $(s_a^o, s_b^o) \in S_a \times S_b$ such that for any internally reciprocal domains (D_{ia}, D_{ib}) with $(s_a^o, s_b^o) \in D_{ia}$, the pair (s_a^o, s_b^o) is an i.c.equilibrium for Γ^i.

Proof. Let us choose a pair $(s_a^o, s_b^o) \in S_a \times S_b$ so that

$$h_a(s_a^o, s_b^o) + h_b(s_a^o, s_b^o) \geq h_a(s_a, s_b) + h_b(s_a, s_b) \text{ for all } (s_a, s_b) \in S_a \times S_b. \tag{6.5}$$

Since $S_a \times S_b$ is a finite set, we can find a pair $(s_a^o, s_b^o) \in S_a \times S_b$ satisfying (6.5).

Since (D_{ia}, D_{ib}) are internally reciprocal, we have, Theorem 3.1.(1), $h_r^{ii}(s_a, s_b) = h_r^{ij}(s_a, s_b) = h_r(s_a, s_b)$ for all $(s_a, s_b) \in \text{Proj}(S_a^i \times S_b^i)$. Hence, using (6.5), we have, for all $s_a \in S_a^i$ and $s_b \in S_b^i$,

$$\begin{aligned} \frac{1}{2}h_a^{ii}(s_a^o, s_b^o) + \frac{1}{2}h_b^{ii}(s_a^o, s_b^o) &\geq \frac{1}{2}h_a^{ii}(s_a, s_b^o) + \frac{1}{2}h_b^{ii}(s_a, s_b^o) \\ \frac{1}{2}h_a^{ij}(s_a^o, s_b^o) + \frac{1}{2}h_b^{ij}(s_a^o, s_b^o) &\geq \frac{1}{2}h_a^{ij}(s_a, s_b^o) + \frac{1}{2}h_b^{ij}(s_a, s_b^o). \end{aligned}$$

The parallel inequalities for the replacement s_b^o by $s_b \in S_b^i$ hold. Hence, (s_a^o, s_b^o) is an i.c.equilibrium in Γ^i. ■

In the above proof, the pair (s_a^o, s_b^o) chosen by (6.5) reaching the maximum payoff sum is independent of person i. Hence, the above proof implies that (s_a^o, s_b^o) is a mutual i.c.equilibrium. We state this fact as a corollary.

Corollary 6.4 (Existence of a Mutual I.C.Equilibrium): Let $(\rho_{ia}, \rho_{ib}) = (\frac{1}{2}, \frac{1}{2})$ for $i = 1, 2$. Then, there is a pair $(s_a^o, s_b^o) \in S_a \times S_b$ such that for any internally reciprocal domain (D_{ia}, D_{ib}) with $(s_a^o, s_b^o) \in D_{ia}$ for $i = 1, 2$, the pair (s_a^o, s_b^o) is a mutual i.c.equilibrium.

In the proof of Theorems 6.3, the pair (s_a^o, s_b^o) is chosen as a global maximization point over the entire matrix. But we should choose one pair (s_a^o, s_b^o) maximizing the simple sum of payoffs over $\mathrm{Proj}(S_a \times S_b)$ centered at this pair. Once this is recognized, a simple algorithm to find such a point is constructed as follows: Take any pair in the matrix. Then, if there is one pair with a higher sum of payoffs obtained by one person's deviation, we move to this pair. If this pair has the same property, then we move again. Then, we will reach one pair without a further improvement. This convergence holds since the matrix is finite and each step has an improvement in the sum of payoffs. The resulting pair may not be a global maximization point.

Next, we will see that an i.c.equilibrium may not exist in the cases where (ρ_{ia}, ρ_{ib}) is twisted enough. To see the nonexistence, we give one lemma.

Lemma 6.5: Let (s_a^o, s_b^o) be an i.c.equilibrium for Γ^i with $(s_a^o, s_b^o) \in D_{ia} \cap D_{ib}$.

(1): If $(s_a, s_b^o) \in D_{ia} \cap D_{ib}$, then $h_a(s_a^o, s_b^o) + h_b(s_a^o, s_b^o) \geq h_a(s_a, s_b^o) + h_b(s_a, s_b^o)$.

(2): If $(s_a^o, s_b) \in D_{ia} \cap D_{ib}$, then $h_a(s_a^o, s_b^o) + h_b(s_a^o, s_b^o) \geq h_a(s_a^o, s_b) + h_b(s_a^o, s_b)$.

Proof. We show only (1). Let $(s_a, s_b^o) \in D_{ia} \cap D_{ib}$. Since (s_a^o, s_b^o) is an i.c.equilibrium, by (5.1), we have

$$\begin{aligned} \rho_{ia} h_a^{ii}(s_a^o, s_b^o) + (1-\rho_{ia}) h_b^{ii}(s_a^o, s_b^o) &\geq \rho_{ia} h_a^{ii}(s_a, s_b^o) + (1-\rho_{ia}) h_b^{ii}(s_a, s_b^o); \quad (6.6) \\ \rho_{ia} h_b^{ij}(s_a^o, s_b^o) + (1-\rho_{ia}) h_a^{ij}(s_a^o, s_b^o) &\geq \rho_{ia} h_b^{ij}(s_a, s_b^o) + (1-\rho_{ia}) h_a^{ij}(s_a, s_b^o). \end{aligned}$$

Since (s_a^o, s_b^o) and (s_a, s_b^o) are in $D_{ia} \cap D_{ib}$, it holds by (3.1) and (3.2) that for $r = a, b$, $h_r^{ii}(s_a^o, s_b^o) = h_r^{ij}(s_a^o, s_b^o) = h_r(s_a^o, s_b^o)$ and $h_r^{ii}(s_a, s_b^o) = h_r^{ij}(s_a, s_b^o) = h_r(s_a, s_b^o)$. Hence, summing up the first and second inequalities of (6.6), we have $h_a(s_a^o, s_b^o) + h_b(s_a^o, s_b^o) \geq h_a(s_a, s_b^o) + h_b(s_a, s_b^o)$. ■

Lemma 6.5 gives necessary conditions for the resulting outcome of an i.c.equilibrium. Using this lemma, we can find the non-existence of an i.c.equilibrium for internally reciprocal domains D_{1a} and D_{1b}.

Example 6.1 (Non-Existence): Consider the game of Table 2.1. Suppose that D_{1a}^{AP} and D_{1b}^{AP} are the active-passive domain given by (2.11) with $(s_a^o, s_b^o) = (\mathbf{s}_{a2}, \mathbf{s}_{b1})$. When $(\rho_{ia}, \rho_{ib}) = (1/2, 1/2)$, $(\mathbf{s}_{a2}, \mathbf{s}_{b1})$ is an i.c.equilibrium and also a weak i.c.equilibrium.

Let $\rho_{ia} = 9/10$ and $\rho_{ib} = 1/10$. In this case, since (D_{1a}, D_{1b}) are internally reciprocal, it follows from Lemma 6.5 that $(\mathbf{s}_{a2}, \mathbf{s}_{b1})$ is a candidate for an i.c.equilibrium. However, we have

$$\begin{aligned} \frac{9}{10} h_a^{ii}(\mathbf{s}_{a1}, \mathbf{s}_{b1}) + \frac{1}{10} h_b^{ii}(\mathbf{s}_{a1}, \mathbf{s}_{b1}) &= 3 > 2.8 \\ &= \frac{9}{10} h_a^{ii}(\mathbf{s}_{a2}, \mathbf{s}_{b1}) + \frac{1}{10} h_b^{ii}(\mathbf{s}_{a2}, \mathbf{s}_{b1}). \end{aligned}$$

Hence, $(\mathbf{s}_{a2}, \mathbf{s}_{b1})$ is not an i.c.equilibrium. The only other candidate with active-passive domains is $(\mathbf{s}_{a1}, \mathbf{s}_{b2})$, but it is not an i.c.equilibrium either.

The pair $(s_a^o, s_b^o) = (\mathbf{s}_{a2}, \mathbf{s}_{b1})$ with the above D_{1a}^{AP} and D_{1b}^{AP} is a weak i.c.equilibrium. On the other hand, many other weak i.c.equilibria appear: Each of $(s_a^o, s_b^o) = (\mathbf{s}_{a2}, \mathbf{s}_{b1})$, $(\mathbf{s}_{a1}, \mathbf{s}_{b2}), (\mathbf{s}_{a1}, \mathbf{s}_{b1}), (\mathbf{s}_{a3}, \mathbf{s}_{b2}), (\mathbf{s}_{a2}, \mathbf{s}_{b3})$ is possibly a weak i.c.equilibrium with the appropriate active-passive domains.

When $\rho_{ia} = 1/3$ and $\rho_{ib} = 2/3$, $(\mathbf{s}_{a2}, \mathbf{s}_{b1})$ becomes again an i.c.equilibrium (and also a weak one, too), though the necessary conditions given by Lemma 6.5 remains valid. This is the fact stated by Theorem 6.2 that with sufficient reciprocity, the payoff-sum maximizer is an i.c.equilibrium.

7. Applications to the Prisoner's Dilemma, Ultimatum Game and Dictator Game

Here, we apply the results of Section 6 to the prisoner's dilemma game, ultimatum game and dictator game. For those games, experimental results differ consistently from the predictions based on the standard equilibrium theory. Cooperative outcomes (equal division) are observed more often in experiments than the predicted non-cooperative outcomes (cf. Cooper *et al.* [4], Güth *et al.* [9], Kahneman *et al.* [13] and also Camerer [3]). Here, we consider some variants of those games, and apply our theory to them.

Prisoner's Dilemma: This is typically expressed as a bimatrix game such as in Table 7.1. Consider the reciprocal active-passive domains $(D_{ia}^{AP}, D_{ib}^{AP})$ with the regular actions (s_a^o, s_b^o) and $\rho_{ia} = 1/2$. It follows from Theorem 6.1 that $(s_a^o, s_b^o) = (\mathbf{s}_{a1}, \mathbf{s}_{b1})$ is the unique i.c.equilibrium. On the other hand, if we multiply the payoffs for role b by 6 to obtain the game of Table 7.2, then $(s_a^o, s_b^o) = (\mathbf{s}_{a1}, \mathbf{s}_{b2})$ becomes the unique i.c.equilibrium. Here we see that the affine transformation of payoffs affects behavioral predictions in our theory. If we go further and change the payoffs to Table 7.3, which maintain the dominant strategies of the game, then the new game has now two i.c.equilibria which are $(\mathbf{s}_{a1}, \mathbf{s}_{b2})$ and $(\mathbf{s}_{a2}, \mathbf{s}_{b1})$.

Table 7.1

	$\mathbf{s}_{b1}$	$\mathbf{s}_{b2}$
$\mathbf{s}_{a1}$	$(5,5)^{IC}$	$(2,6)$
$\mathbf{s}_{a2}$	$(6,2)$	$(3,3)^{NE}$

Table 7.2

	$\mathbf{s}_{b1}$	$\mathbf{s}_{b2}$
$\mathbf{s}_{a1}$	$(5,5)$	$(2,36)^{IC}$
$\mathbf{s}_{a2}$	$(6,12)$	$(3,18)^{NE}$

Table 7.3

	$\mathbf{s}_{b1}$	$\mathbf{s}_{b2}$
$\mathbf{s}_{a1}$	$(5,5)$	$(2,10)^{IC}$
$\mathbf{s}_{a2}$	$(10,2)^{IC}$	$(3,3)^{NE}$

Contrary to these results, an i.c.equilibrium becomes very different in the case of the non-reciprocal active domains. In each of the above table, it follows from Theorem 5.2.(2) that $(s_a^o, s_b^o) = (\mathbf{s}_{a2}, \mathbf{s}_{b2})$ remains the unique i.c.equilibrium, which is also (a dominant strategy) Nash equilibrium. We need several comments on our predictions.

First, the above three bimatrix games are all regarded as the prisoner's dilemma from the standard game theoretical point of view. However, the i.c.equilibrium concept behaves differently in those games. In the full reciprocal case, the i.c.equilibrium moves to the payoff-sum maximization points in those games. This is one possible prediction of our theory, which can be tested in experiments.

Second, in the non-reciprocal case, the i.c.equilibrium coincides with the Nash equilibrium as was stated in Theorem 5.2.(2). The reader may wonder whether this is consistent with the existing experimental results, which state that the cooperative outcome more likely results. However, one important difference we should notice is that in our theory, the payoff functions are assumed to be *a priori* not known. If this "not known" is taken properly into account in experiments, we expect that $(s_a^o, s_b^o) = (\mathbf{s}_{a2}, \mathbf{s}_{b2})$ could more likely result.

Finally, the reader may also wonder why the cooperative outcome has been observed in experiments when the payoffs are assumed to be known to the subjects. As far as the game is symmetric such as in Table 7.1, this "known" assumption works as a substitute for role-switching in providing information about the payoff of the other role, and the cooperative outcome may observed without role switching. However, in non-symmetric cases such as in Table 7.2, we do not expect the same results. These cases need further experimental study as well as for an extension of our theory itself to incorporate a postulate different from Ob2 (observation only of his own payoff pair).

Ultimatum Game: Suppose that the 2-role game is given as follows: A person assigned to role a proposes a division of \$100 to persons 1 and 2, and a person assigned to role b receives the proposal (x_a, x_b) and chooses an answer Y or N to the proposal. We assume that only three alternative choices are available at a, i.e., $S_a = \{(99, 1), (50, 50), (1, 99)\}$. The person at role b chooses Y or N contingent upon the offer made by a, i.e., $S_b = \{(\alpha_1, \alpha_2, \alpha_3) : \alpha_1, \alpha_2, \alpha_3 \in \{Y, N\}\}$. If the person at role a chooses $(99, 1)$ and the person at b chooses $(\alpha_1, \alpha_2, \alpha_3)$, then the outcome depends only upon α_1; if $\alpha_1 = Y$, then they receive $(99, 1)$ and if $\alpha_1 = N$, then they receive $(0, 0)$. For the other cases, we define payoffs in a parallel manner. The game is depicted in Fig.7.1.

This game has a unique backward induction solution: $((99, 1), (Y, Y, Y))$.This is quite incompatible with experimental results, which have indicated that $(50, 50)$ is more likely chosen by the mover at a, as mentioned above.

We assume one additional component for the persons. They have a *strictly* concave and monotone utility function $u(m)$ over $[0, 100]$. This introduction does not change the above equilibrium outcome. But it changes the i.c.equilibrium drastically.

Under the assumption that person i has the reciprocal active-passive domains $D_{ia}^{AP} =$

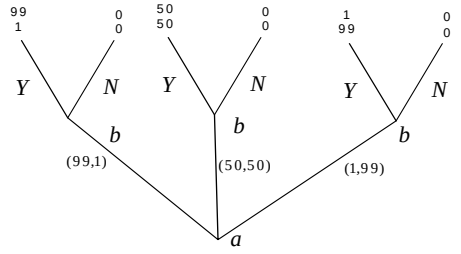

Figure 7.1: Ultimatum Game

D_{ib}^{AP} and $\rho_{1a} = 1/2$, a pair $((99,1),(Y,Y,Y))$ is not an i.c.equilibrium since

$$\begin{aligned} & \frac{1}{2}h_a^{ii}((99,1),(Y,Y,Y)) + \frac{1}{2}h_b^{ii}((99,1),(Y,Y,Y)) \\ = \; & \frac{1}{2}u(99) + \frac{1}{2}u(1) < u(50) = \frac{1}{2}u(50) + \frac{1}{2}u(50) \\ = \; & \frac{1}{2}h_a^{ii}((50,50),(Y,Y,Y)) + \frac{1}{2}h_b^{ii}((50,50),(Y,Y,Y)). \end{aligned}$$

The inequality follows the strict concavity of u. In this game, an i.c.equilibrium is given as $((50,50),(\alpha_1,Y,\alpha_3))$, where α_1, α_3 are not determined. We find that the concept of i.c.equilibrium is consistent with the experimental results. In fact, we have other i.c.equilibria, e.g., $((99,1),(Y,N,N))$ and even $((1,99),(N,N,Y))$, which are also Nash equilibria of this game.

There are several issues here. One is that we have treated this game as a strategic game to fit into the theory given in this paper. In order to study it as an extensive game, we need to extend our theory to extensive games or information protocols such as in [16] and [17]. Another issue is that the i.c.equilibrium does not consider joint deviations, so equilibria like $((99,1),(Y,N,N))$ can persist. As mentioned in Section 4.2, we could have extended our theory to include joint deviations, but chose not to do so, since it would include other conceptual problems. With such an extension, we could discuss how the other equilibria such as $((99,1),(Y,N,N))$ may or may not remain in our theory.

Here, instead of extending the present theory, we simplify the ultimatum game so as to treat it as a strategic game to show how the results become clear cut. We will treat a simpler version of the dictator game given by Kahneman *et al.* [13] (see also Camerer [3] for a survey of experimental studies of dictator games).

Dictator Game: Let us eliminate action N from each move of role b. The game is depicted as Fig.7.2. This has no action choice for role b, and thus, it is regarded a 1-role

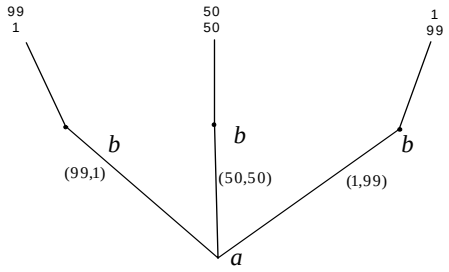

Figure 7.2: Dictator Game

game from the standard game theoretic point of view. However, payoffs to role b matter in our theory. First, we consider:

Case 1: Reciprocal Active-Passive Domains: Here, we first specify the domain and frequencies of role-switching with $(s_a^o, s_b^o) = ((50, 50), Y)$:

$$\begin{aligned} D_{ia} &= \{((99,1),Y),((50,50),Y),((1,99),Y)\}, D_{ib} = \{((50,50),Y)\} \qquad (7.1) \\ \rho_{1a} &= 1/2. \end{aligned}$$

Then $S_a^i = \{(99,1),(50,50),(1,99)\}$ and $S_b^i = \{Y\}$. Here, we have the unique i.c.equilibrium $((50,50),Y)$. Indeed, when they switch the roles, one person obtains \$99 and \$1 with frequencies $1/2$ and $1/2$, respectively. This alternating payoffs are less preferred to taking \$50 constantly, since the utility function u is strictly concave.

To discuss whether this result can be regarded as capturing the experimental results reported so far, we consider another extreme case.

Case 2: Non-reciprocal Domains:

$$\begin{aligned} D_{1a} &= \{((99,1),Y),((50,50),Y),((1,99),Y)\} \text{ and } D_{1b} = \emptyset \qquad (7.2) \\ \rho_{1a} &= 1. \end{aligned}$$

That is, person 1 always chooses a division of \$100, and person 2 follows it. In this case, we have also a unique i.c.equilibrium $((99,1),Y)$: Person 1 exclusively enjoys role a. In this case, the domains for person 2 are: $D_{2a} = \emptyset$ and $D_{2b} = \{((99,1),Y)\}$.

The results for the above two cases are extremely opposite. We should discuss whether the prediction of our theory may reconcile the discrepancy between the game theory and reported experimental results.

Discussions of the Above Results: Social Contexts:

A lot of experimental studies are reported based on the prisoner's dilemma, ultimatum game, and dictator game. As already stated, the experimental results consistently

differ from the non-cooperative game-theoretical predictions. The results are rather closer to our cooperative results. However, experimental theorists have tried to interpret their results in terms of "fairness", "altruism", and/or "social preferences", which are expressed as constraint maximization of additional objective functions (cf., Camerer [3]). In contrast, we have extended and/or specified the basic social context, and derived the emergence of cooperation. Thus, our treatment is very different from what have been discussed in the literature of behavioral economics and game theory. Perhaps, ours will serve a new theoretical viewpoint to experimental economics.

We have already discussed about possible experimental studies for the prisoner's dilemma. Here, we discuss only the dictator game and our result. One possible hypothesis is that the fully reciprocal Case 1 with equal sharing corresponds to the standard experimental design where the roles and the opponents are chosen randomly in each round keeping their anonymity. This experimental design already captures our internal reciprocity well and the experimental results of sharing fit well.

Exactly speaking, we find a gap between the above experimental design and our internal reciprocity, since the random choice of a subject from the pool differs from the role-switching of the two fixed subjects. Nevertheless, our entire view explains this gap: A basic assumption of inductive game theory is that a person takes patterned behavior in a complex social web, meaning that he behaves in the same or similar situation following the same pattern of behavior. The situation our theory targets is repeated but it may be scattered in the social web, like Fig.1.1. A subject taken from society brings his behavior pattern and behaves following it in an experiment. The cooperative behavior described by an i.c.equilibrium may be taken by a person to an experiment where he again behaves cooperatively.

Some alternative experimental design may be developed to capture the non-reciprocal Case 2. In this design, the roles could be fixed over some rounds, say 20 rounds, with an anonymous opponent. Here, we might expect the non-sharing i.c.equilibrium of case 2 to result.

The idea of patterned behavior should be applied even to optimization behavior. Though we have described the optimal behavior of a person as an i.c.equilibrium, this does not imply that a subject is an instantaneous optimizer. Rather, each typically follows his patterned behavior and only sometimes maximizes his payoffs. Optimization results only in the long-run. This idea is an answer from the entire approach of inductive game theory to the question: "*How do socio-cognitive dimensions influence behavior in games?*" in Camerer [3], p.476.

Now we turn to morality or fairness. It is our contention that as far as a situation is recurrent and reciprocal enough, the persons possibly cooperate in the form of the simple payoff sum maximization. Since this is, perhaps, quite pervasive for human relations among small numbers of people, they could have such patterned behavior, and consequently, such behavior is then observed in experiments. This gives an "anthropo-

logical", i.e., "experiential" grounding for morality. This differs from the rationalistic school of morality - - it comes from rationalistic reasoning about morality (such as in Harsanyi [10]). It also differs from Adam Smith's [27] "moral sentiments" - - people are born with a moral sense. In our case, the "morality" of the form of the payoff-sum maximization emerges from social interactions and role-switching in complex social web, and is neither rationalized nor inborn. We regard this as an anthropological foundation for the "utilitarianism" expressed in the form of Theorem 6.1.

8. Externally Reciprocal Relations

Our primary concern was what happens with experiences in the mind of one person. Actually, since people are in a game setting, experiences and understandings from them are also externally interactive and affect each other. In this section, we will consider various external reciprocal relations. Our basic idea is that the persons' reciprocal relationships are gradually emerging as time is going on. In this process, an active experience and a passive experience may behave quite differently. In this section, we will focus on unilateral trials and the generation of a resulting memory kit based on such trials.

The starting point is as follows. Suppose that persons 1 and 2 have their accumulated domains $D_1 = (D_{1a}, D_{1b})$ and $D_2 = (D_{2a}, D_{2b})$, respectively, with the regular actions (s^o_a, s^o_b). These accumulated domains should be correlated since the passive experiences of one person are generated by active experiences of the other. Using this idea, we could impose the following condition on domains of accumulation:

Active generates Passive: for all $s_r \in S_r$, $r = a, b$ and $i, j = 1, 2$ $(i \neq j)$,

$$(s_r; s^o_{-r}) \in D_{j(-r)} \text{ implies } (s_r; s^o_{-r}) \in D_{ir}. \tag{8.1}$$

That is, if person j has a passive experience, then person i must have this as an active experience causing j's passive experience. This is of the same nature as the Postulates EP3 and EP4 of Section 2. Based on these postulates, (8.1) formulates the idea that a person is more sensitive to being active with respect to memories. This gives an element of reciprocity but is only a necessary form of reciprocity. For example, the non-reciprocal active domains D^N_1 and D^N_2 given by (2.9) still satisfy (8.1).

As time is going on, each person may have learned also passive experiences. Eventually, the converse of (8.1) could hold:

Equal Sensitivity of Active/Passive Experiences: for all $s_r \in S_r$, $r = a, b$ and $i, j = 1, 2$ $(i \neq j)$,

$$(s_r; s^o_{-r}) \in D_{j(-r)} \text{ if and only if } (s_r; s^o_{-r}) \in D_{ir}. \tag{8.2}$$

The non-reciprocal active domains D_1^N and D_2^N no longer satisfy this condition. If we keep the assumption that they do not switch roles but if (8.2) is assumed, then we should amend the non-reciprocal active-passive domains D_1^{NAP} and D_2^{NAP} described in (1) of Section 2.2. We can see that the amendments do not change the behavioral consequence from Theorem 4.1, though the tp-understanding g^{12} changes slightly, i.e., person 1 may now recognize a larger action set S_b^i.

Role-Switching with Similar Frequencies: The above example suggests that (8.2) is not enough to establish external reciprocal relationships between 1 and 2. We need also the assumption that they switch the roles from time to time with relatively equal frequencies.

Nevertheless, the equal sensitivity (8.2) and the frequency-wise reciprocity are still not enough for the fully reciprocal relationships.

Example. 8.1 (Different Trials): Consider the game in Table 2.1 and the following D_1, D_2 with the regular actions $(s_a^o, s_b^o) = (\mathbf{s}_{a1}, \mathbf{s}_{b1})$;

$$\begin{aligned} D_{1a} &= \{(\mathbf{s}_{a1}, \mathbf{s}_{b1}), (\mathbf{s}_{a2}, \mathbf{s}_{b1}), (\mathbf{s}_{a1}, \mathbf{s}_{b3})\}, \text{ and } D_{1b} = \{(\mathbf{s}_{a1}, \mathbf{s}_{b1}), (\mathbf{s}_{a1}, \mathbf{s}_{b2}), (\mathbf{s}_{a3}, \mathbf{s}_{b1})\}; \\ D_{2a} &= \{(\mathbf{s}_{a1}, \mathbf{s}_{b1}), (\mathbf{s}_{a3}, \mathbf{s}_{b1}), (\mathbf{s}_{a1}, \mathbf{s}_{b2})\}, \text{ and } D_{2b} = \{(\mathbf{s}_{a1}, \mathbf{s}_{b1}), (\mathbf{s}_{a1}, \mathbf{s}_{b3}), (\mathbf{s}_{a2}, \mathbf{s}_{b1})\}. \end{aligned}$$

That is, person 1 makes trials of only the second actions $\mathbf{s}_{a2}$ at role a and $\mathbf{s}_{b2}$ at role b, while person 2 makes trials of only the third actions $\mathbf{s}_{a3}$ at a and $\mathbf{s}_{b3}$ at b. Even though (8.2) holds, and their roles are switched, their differences in trial behaviors generates different domains of experiences, i.e., $D_{ia} \neq D_{ib}$ for $i = 1, 2$ and $D_{1r} \neq D_{2r}$ for $r = a, b$, though $D_{1a} = D_{2b}$ and $D_{2a} = D_{1b}$. These nonequivalences prevent them from constructing meaningful tp-understandings.

Thus, we need to take one more step to obtain full reciprocity

The Same Trials: The two persons switch the roles and make similar trials as well. The extreme case is formulated as: for all $s_r \in S_r$, $r = a, b$ and $i, j = 1, 2$ $(i \neq j)$,

$$(s_r; s_{-r}^o) \in D_{ir} \text{ if and only if } (s_r; s_{-r}^o) \in D_{jr}. \tag{8.3}$$

That is, they make the same trials at each role.

We can change (8.2) and (8.3) to equivalent but mathematically clearer conditions.

Lemma 8.1 (Internal-External Reciprocity). Conditions (8.2) and (8.3) hold for (D_{1a}, D_{1b}) and (D_{2a}, D_{2b}) if and only if

(1)(Internal Reciprocity): $\text{Proj}(D_{ia}) = \text{Proj}(D_{ib})$ for $i = 1, 2$;

(2)(External Reciprocity): $\text{Proj}(D_{1r}) = \text{Proj}(D_{2r})$ for $r = a, b$.

Proof. When (1) and (2) hold, the four sets, $\text{Proj}(D_{ir})$, $i = 1, 2$ and $r = a, b$ coincide.

Hence, the if-part is straightforward. We prove the only-if part. Suppose (8.2) and (8.3) for (D_{1a}, D_{1b}) and (D_{2a}, D_{2b}).

Consider (1). Let $(s_a, s_b) \in \mathrm{Proj}(D_{1a})$. This means that $(s_a, s_b) = (s_a, s_b^o)$ or (s_a^o, s_b). Fist, let $(s_a, s_b) = (s_a, s_b^o)$. Then, $(s_a, s_b^o) \in \mathrm{Proj}(D_{2a})$ by (8.3), which is written as $(s_a; s_{-a}^o) \in \mathrm{Proj}(D_{2a})$. By (8.2), we have $(s_a; s_{-a}^o) \in \mathrm{Proj}(D_{1(-a)})$, i.e., $(s_a, s_b^o) \in \mathrm{Proj}(D_{1b})$. Next, let $(s_a, s_b) = (s_a^o, s_b)$. Thus, $(s_b; s_{-b}^o) \in \mathrm{Proj}(D_{1(-b)})$. We have $(s_b; s_{-b}^o) \in \mathrm{Proj}(D_{2b})$ by (8.2). Hence, by (8.3), we have $(s_b; s_{-b}^o) \in \mathrm{Proj}(D_{1b})$. We have shown $\mathrm{Proj}(D_{ia}) \subseteq \mathrm{Proj}(D_{ib})$. The converse can be obtained by a symmetric argument. Thus, we have (1).

Consider (2). Let $(s_a, s_b) \in \mathrm{Proj}(D_{1a})$. This means that $(s_a, s_b) = (s_a, s_b^o)$ or (s_a^o, s_b). Let $(s_a, s_b) = (s_a, s_b^o)$. By (8.3), we have $(s_a, s_b^o) \in \mathrm{Proj}(D_{2a})$. i.e., $(s_a; s_{-a}^o) \in \mathrm{Proj}(D_{2a})$. Now, let $(s_a, s_b) = (s_a^o, s_b)$. By (1), $(s_a^o, s_b) \in \mathrm{Proj}(D_{1b})$. This is written as $(s_b; s_{-b}^o) \in \mathrm{Proj}(D_{1b})$. By (8.2), we have $(s_b; s_{-b}^o) \in \mathrm{Proj}(D_{2a})$. We have shown that $\mathrm{Proj}(D_{1a}) \subseteq \mathrm{Proj}(D_{2a})$. The converse can be obtained by a symmetric argument. Thus we have (2). ■

Hence, when (8.2) and (8.3) hold, these $\mathrm{Proj}(D_{ir})$ coincide for $i = 1, 2$ and $r = a, b$. Hence, as far as the frequency weights are reciprocal, i.e., $\rho_{1a} = \rho_{2a} = 1/2$, an i.c.equilibrium and a mutual i.c.equilibrium support an cooperative outcome up to the experienced actions.

In Theorem 3.2, we have already seen that internal reciprocity (1) is necessary and sufficient for g^{ii} and g^{ij} to coincide within the mind of person i. The next step is to consider when the two persons reach the same views. In this case, under the assumption of $\rho_{1a} = \rho_{2a} = 1/2$, a mutual i.c.equilibrium makes sense.

Actually, (8.2) and (8.3) are necessary and sufficient for all g^{ii} and g^{ij} $(i, j = 1, 2, i \neq j)$ to coincide across persons. We state this result as a theorem.

Theorem 8.2.(Internally and Externally Reciprocal Relations): (8.2) and (8.3) hold for (D_{1a}, D_{1b}) and (D_{2a}, D_{2b}) if and only if for any $r = a, b$ and $i, j = 1, 2$ $(i \neq j)$,
(1): $S_r^i = S_r^j$;
(2): for any $(s_a, s_b) \in \mathrm{Proj}(S_a^1 \times S_b^1)$ and θ_a, θ_b, $h_r^{ii}(s_a, s_b) = h_r^{ij}(s_a, s_b) = h_r(s_a, s_b)$.

Proof. **(Only-If)**: Suppose that (8.2) and (8.3) hold for (D_{1a}, D_{1b}) and (D_{2a}, D_{2b}). Then, Lemma 8.1 states that $\mathrm{Proj}(D_{ir})$'s all coincide. Hence, (1) is satisfied by the d-understanding $g^{ii} = (a, b, S_a^i, S_b^i, h_a^{ii}, h_b^{ii})$ and tp-understanding $g^{ij} = (a, b, S_a^i, S_b^i, h_a^{ij}, h_b^{ij})$ for $i, j = 1, 2$ $(i \neq j)$. Assertion (2) also follows by (3.1) and (3.2).

(If): By (1) and (2.3), we have, for $i = 1, 2$, $\mathrm{Proj}(D_{ia} \cup D_{ib}) = \mathrm{Proj}(S_a^1 \times S_b^1)$. Let $(s_a, s_b) \in \mathrm{Proj}(D_{ia} \cup D_{ib})$. Then, since $h_r^{ii}(s_a, s_b) = h_r^{ij}(s_a, s_b) = h_r(s_a, s_b)$ for any θ_a, θ_b by (2), we have $(s_a, s_b) \in D_{ir} \cap D_{i(-r)}$. This holds for $i = 1, 2$. Hence, $(s_a, s_b) \in \mathrm{Proj}(D_{ia})$ and $(s_a, s_b) \in \mathrm{Proj}(D_{ib})$. Hence, we have shown (1) and (2) of Lemma 8.1. Thus, (8.2) and (8.3) hold for (D_{1a}, D_{1b}) and (D_{2a}, D_{2b}). ■

An implication of Theorem 8.2 is that under (8.2), (8.3) and the frequency assumption that $(\rho_{ia}, \rho_{ib}) = (\frac{1}{2}, \frac{1}{2})$ for $i = 1, 2$, person i can predict correctly the other person's weighted payoff function over the relevant domains, that is, $H^{ij}([s_r; s^o_{-r}]_a, [s_r; s^o_{-r}]_b) = H^{jj}([s_r; s^o_{-r}]_a, [s_r; s^o_{-r}]_b)$ for any $s_r \in S^j_r = S^i_r$. Hence, those persons think about the game in the perfectly synchronized manner, *a fortiori*, if (s^o_a, s^o_b) is a mutual i.c.equilibrium, then they reach the understanding that it is an i.c.equilibrium for both persons.

Although we have explore the possibility of these two types of reciprocities to be reached after the situation has been played with full role-switching, we do not claim that this happens necessarily. Here, we give only one example where each has internally reciprocal domains but they are not externally reciprocal

Example 8.2 (Internally Reciprocal for each but not External Reciprocal). Two persons 1 and 2 have played the game with full role-switching, and have made the same trial deviations from the regular actions. Now, suppose that person 1 has a stronger memory ability than person 2. In this case, person 1 keeps more experiences than 2, while internal reciprocity holds for each person, i.e., $\text{Proj}(D_{1a}) = \text{Proj}(D_{1b}) \supsetneq \text{Proj}(D_{2a}) = \text{Proj}(D_{2b})$. In this case, each person has the same d- and tp-understandings, but they are different over the persons.

In this case, the dynamics suggested in Fig.4.1 may not work externally. For example, person 1 thinks that a deviation s_a gives a better weighted payoff, and he thinks that person 2 thinks in the same manner. But, if the experience (s_a, s^o_b) is not accumulated in person 2's mind, person 2 does not deviate as 1 predicts. In this case, person 1 may find that person 2's i.d.view is different.

This kind of a difference in their views may be a source for their communications. This is beyond the scope of this paper and will be discussed in a separate paper.

9. Conclusions

We have introduced the concept of social roles into inductive game theory, and have given an experiential foundation of the other's beliefs. Based on this foundation, we have shown the possibility for the emergence of cooperation and argued that persons are more likely to cooperate when their role-switching is more reciprocal. In this section, we first summarize our findings in this paper, and next we discuss extensions and future work.

9.1. Summary of Findings

It was our basic postulate that a person's understanding of the other's thinking should be experiential. We introduced role-switching so that person i could experience and obtain an experiential understanding of the other's thinking.

In our exploration of a person's transpersonal understanding of the other, we have taken several steps exemplified by various postulates. We postulated in TP1 (projection of self) and TP2 (experiential reason to believe) that each person projects his own experiences onto the other provided he has experiential reason to believe the other has had the same experience. These postulates were summarized in the requirement that both $(s_a, s_b) \in D_{ir}$ and $(s_a, s_b) \in D_{ir}$ in the definition of $h_r^{ij}(s_a, s_b)$ in (3.2). This will be discussed below more.

The transpersonal understanding of the other's thinking requires reciprocity in role-switching. With such reciprocity, it became natural to consider the frequency weighted payoff of a person across roles. Correspondingly, we developed the concept of an i.c.equilibrium within such a framework.

Nevertheless, with different degrees of reciprocity, we have many cases for the domains of accumulation generated, some of which were well suited for cooperation, and others not. The reciprocal active-passive domain was shown to be well suited for cooperation to emerge in an i.c.equilibrium (Theorem 6.1). On the other hand, the non-reciprocal and reciprocal active domains generate only non-cooperative Nash equilibria as i.c.equilibria (Theorem 5.2). In this paper, we pursued only a few cases to express the main thrust of the arguments about the potential for the emergence of cooperation.

In Section 7, we discussed the coherency of our theory with experimental results from the prisoner's dilemma, ultimatum, and dictator games. We also proposed some alternative experimental designs to test the relevance of role-switching for behavior in experiments, which will serve a connection to experimental/behavioral economics/game theory (cf., Camerer [3]).

Section 8 gave external conditions for the internal reciprocity which was at the heart of the emergence of cooperation in our theory. This exploration showed that in addition to reciprocal role-switching, the same trials by both persons, and equal and broad sensitivities were sufficient (Theorem 8.2) to generate the equivalent understandings that are fertile grounds for cooperation.

9.2. Extensions and Future Work

First, we discuss some implicit assumptions underlying of formulation of person i's derivation of the tp-understanding about the other's understanding. It is experiential in the sense that all components are derived from his own accumulated experiences. Here, we need social roles, role-switching, and also the basic assumption that the 2-role game is given independent of persons (actors). In this sense, we have followed the tradition of symbolic interactionism from Mead [23]. In reality, our treatment is an idealization. Not only this, we need other assumptions. These were specified from place to place in this paper.

Specifically, the definition of person i's tp-understanding g^{ij} from his memory kit κ_i

includes such an assumption. The salient part is the condition $(s_a, s_b) \in D_{i(-r)}$ in the definition of $h_r^{ij}(s_a, s_b)$ in (3.2). This means that person i has the experience (s_a, s_b) in his domain $D_{i(-r)}$: He infers that person j must have also this experience, and project his experienced payoff $h_r(s_a, s_b)$ onto j. That is, (s_a, s_b) must be a common experience for persons i and j from the viewpoint of person i. This sounds like the requirement of some evidence for the definition of common knowledge in Lewis [21]. This will possibly serve a bridge to epistemic logic; in particular, to common knowledge logic of Fagin *et al.* [7] (see also Kaneko [14]).

In our context, if a person experiences one pair (s_a, s_b) from both roles, he would infer/guess that the other person has the same experience from both roles, and also that the other infers the symmetric statement. If we pursue, rigorously, this argument as an infinite regress, then we would have common knowledge (beliefs) in the sense of an infinite hierarchy of beliefs (see Kaneko [15], Chap.4 for this argument of an infinite regress). In this case, we need other assumptions on an hierarchy of logical abilities of the persons. As Lewis [21] did not intend to mean an infinite hierarchy of knowledge (beliefs), it would be better to stop at some shallow interpersonal depths of nested beliefs. We will discuss a rigorous treatment of this in the epistemic logic of shallow depths (Kaneko-Suzuki [20]) in a separate paper.

Next, we turn to some extensions like the emergence of cooperation in n-role games. Notice that emergence of cooperation is conditional upon the degree of reciprocity of role-switching. We restricted ourselves to a 2-person situation and still cooperation needs a specific reciprocity. Therefore, our result may be interpreted as showing a difficulty in reaching cooperation. One immediate question is to ask what would happen with the present study in a 3- or more persons case. This remains an open problem, but we should give our thought about it.

In an n-person case, since one person can experience a few social roles only, it might not be appropriate to extend directly the result of this paper into the n-person case. Rather we should consider possibilities of cooperations of 2- or 3-person groups in the entire n-person game. These groups of small sizes could represent the extent of a person's cooperation potential. In this limited sense, our research does not suggest us to return to the standard n-person cooperative game theory from von Neumann-Morgenstern [28].

Rather, patterned behavior in different but similar situations may be a key to have an extension of our theory. This is related to the basic presumption of inductive game theory: A social situation formulated as a 2-role game (more generally, an n-role game) is not isolated from other social situations in the entire social web. As a research strategy in this paper, we focused on a specific 2-role game, but we should not forget that this simple case belongs to the complex social web depicted as Fig.1.1. Overlaps and connections between similar situations becomes unavoidable.

Also, we remind the readers that our behavioral postulate is of patterned (regu-

lar) behavior, rather than instantaneous payoff maximization. This patterned behavior may have some uniformity (regularity), which could ease some difficulty in reaching cooperation such as one difficulty of multiplicity we met in the ultimatum game in Section 7. This thought may suggest more experiential studies of behavior in society and experimental studies in labs.

References

[1] Akiyama, E., R. Ishikawa, M. Kaneko and J. J. Kline, (2008), A Simulation Study of Learning a Structure: Mike's Bike Commuting, to appear in *Economic Theory*.

[2] Axelrod (1984), *The Evolution of Cooperation*, Basic Books, New York.

[3] Camerer, C., (2003), *Behavioral Game Theory*, Princeton University Press, Princeton.

[4] Cooper, R., D. V. DeJong, R. Forsyth, and T. W. Ross (1996), Cooperation without Reputation: Experimental Evidence from Prisoners' Dilemma Games, *Games and Economic Behavior* 12, 187-218.

[5] Collins, R. (1988), *Theoretical Sociology*, Harcourt Brace Javanovic, New York.

[6] Cooley, C. H., (1902), *Human Nature and the Social Order*, Scribner, New York.

[7] Fagin, R., J.Y. Halpern, Y. Moses and M. Y. Vardi, (1995), *Reasoning about Knowledge*, The MIT Press, Cambridge.

[8] Gardenfors, P., (2008) The Role of Intersubjectivity in Animal and Human Cooperation, *Biological Theory* 3, 51-62.

[9] Güth, W., Schmittberger, Schwarze, (1982), An Experimental Analysis of Ultimatum Bargaining, *Journal of Economic Behavior and Organization* 3, 367-388.

[10] Harsanyi, J. C., (1953), Cardinal utility in welfare economics and in the theory of risk-taking, *Journal of Political Economy* 61, 434-435.

[11] Hart, S., (2006), Robert Aumann's Game and Economic Theory, *Scandinavian Journal of Economics* 108, 185-211.

[12] Hu, T.-W., (2008), Expected Utility Theory form the Frequentist Perspective, to appear in *Economic Theory*.

[13] Kahneman, D., J. L. Knetsch and R. Thaler, (1986), Fairness as a Constraint on Profit Seeking: Entitlements in the Market, *American Economic Review* 76, 728-741.

[14] Kaneko, M., (2002), Epistemic logics and their game theoretical applications: Introduction. *Economic Theory* 19, 7-62.

[15] Kaneko, M., (2004), Game Theory and Mutual Misunderstanding, Springer, Berlin.

[16] Kaneko, M., and J. J. Kline, (2008a), Inductive Game Theory: a Basic Scenario, *Journal of Mathematical Economics* 44, 1332-1363.

[17] Kaneko, M., and J. J. Kline, (2008b), Information Protocols and Extensive Games in Inductive Game Theory, *Game Theory and Applications 13,* 57-83.

[18] Kaneko, M., and J. J. Kline, (2008c), Partial Memories, Inductively Derived Views, and their Interactions with Behavior, to appear in *Economic Theory.*

[19] Kaneko, M., and A. Matsui, (1999), Inductive Game Theory: Discrimination and Prejudices, *Journal of Public Economic Theory* 1, 101-137. Errata: the same journal 3 (2001), 347.

[20] Kaneko, M., and N.-Y. Suzuki, (2002), Bounded interpersonal inferences and decision making, *Economic Theory* 19, 63-103.

[21] Lewis, D. (1969), *Convention: A Philosophical Study,* Harvard University Press, Cambridge.

[22] Matsui, A., (2008), A Theory of Man as a Creator of the World, *Japanese Economic Review* 59, 19-32.

[23] Mead, G. H., (1934), *Minds, Self and Society*, Chicago University Press, Chicago.

[24] Mendelson, E., (1987), *Introduction to mathematical logic.* Monterey: Wadsworth.

[25] Nash, J. F., (1951), Noncooperative Games, *Annals of Mathematics* 54, 286-295.

[26] Nash, J. F., (1953), Two-person Cooperative Games, *Econometrica* 21, 128-140.

[27] Smith, A., (1759, 2007), *The Theory of Moral Sentiments*, Cosimo Classics, London.

[28] von Neumann, J., and O. Morgenstern, (1944), *Theory of Games and Economic Behavior*, Princeton University Press, Princeton.