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

Mamoru Kaneko‡ and J. Jude Kline §

26 April 2009; this version 28 February 2012

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 with occasional role-switching. Each person accumulates experiences of both roles by switching roles, and these experiences become the source for 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 views for his behavior revision, and we define an equilibrium called an \textit{intrapersonal coordination equilibrium}. Based on this concept, we show that cooperation will emerge as the degree of reciprocity increases.

Key Words: Inductive Game Theory, Role-Switching, Reciprocity, Strategic Game

JEL Classification Numbers: C70, D80

1. Introduction

The problem of how a person obtains his own beliefs about other persons’ thoughts has not yet been adequately addressed in the game theory and economics literature. Instead, it is typical to assume well-formed beliefs about the game for each player. The

∗The authors are partially supported by Grant-in-Aids for Scientific Research No.21243016, Ministry of Education, Science and Culture, and Australian Research Council Discovery Grant DP0560034.
†The authors thank Burkhard Schipper and Yusuke Narita for detailed comments on an earlier version of the paper.
‡University of Tsukuba, Ibaraki 305-8573, Japan (kaneko@shako.sk.tsukuba.ac.jp)
§University of Queensland, Brisbane, QLD 4072, Australia, (j.kline@uq.edu.au)
beliefs we refer to are beliefs about the structure of the game including such aspects as the relevant players, possible sequence of moves, available actions at each move, and resulting outcomes. These are not simply probabilistic beliefs about chance or strategies of the other players. The present authors [9], [10] and [11] have developed inductive game theory (IGT)\(^1\) in order to explore experiential sources for individual beliefs.

This paper takes one step further by positing that role-switching acts as an experiential source for one’s beliefs about the beliefs of others. By occasional role-switching, each person obtains a richer set of experiences, and he may use it to construct a richer social view. In addition, role-switching may have some behavioral implications on cooperation among involved people. In this paper, we consider only 2 person role-switching situations, since there is already much to be learned from them.

A simple example of 2 person role-switching as a source for experiential learning about the other is found in the daily activities between a wife and a husband. They may divide their family tasks into the roles of “raising the children” and “budget allocation”. By switching these roles from time to time, each may learn the other’s perspective, and ultimately, the partners may find a more cooperative approach to their family affairs. Although we will study a specific 2 person situation with role-switching, it important to emphasize that each such situation occurs within an entire social web like the one described in Fig.1.1.

Let us look at this figure in more detail in order to elaborate on the development of our theory. First, \(G^0(1,2)\) describes an instance of a 2 person game \(G^0\) with role \(a\) taken by person 1 and role \(b\) taken by person 2. The instance \(G^0(2,1)\) is based on the same game \(G^0\), except the roles are switched. As mentioned above, we have included different games like \(G^2\) to keep in mind that each player participates in a variety of games, not only \(G^0\).

In order to describe the role-switching, we distinguish between a role and a person. A role corresponds to what game theorist refer to as a player. A person, on the other hand, is an individual who participates in a social web like Fig.1.1 taking on various roles in various game. For our analysis of role-switching, we will focus on a specific pair of people, 1 and 2, and a specific 2-role game \(G^0\) with roles \(a\) and \(b\). We presume that each person \(i = 1, 2\) can separate this \(G^0\) from the other 2 person games, and he keeps memories, from playing \(G^0\) with the other person, in the form of a memory kit\(^2\) described in Section 2. We remind the reader that in IGT a person has little prior information about \(G^0\). He uses his memories over the repeated plays of \(G^0\) to both construct his view of \(G^0\), and also to adjust his behave in \(G^0\). Our interest in how role-switching affects his view and behavior.

\(^1\)A seminal form of IGT was given in Kaneko-Matsui [12].

\(^2\)The cognitive limititions on a person are implicit in our formulation of a memory kit which ignores the precise sequence of past plays of \(G^0\). This formulation is justified by the epistemic postulates formulated in Section 2.2 which embody bounded rationality aspects of people.
Figure 1.1: Social Web

Our development and findings are greatly influenced by the works of Mead [15] (cf., Collins [4], Chap.7) on the importance of role-switching for obtaining a social view, and Lewis [14] on common knowledge. When two people switch social roles reciprocally, each has played each role many times and has seen the other person in each corresponding role. Based on these experiences, each may guess that the other’s beliefs and perspective are similar to his own. This reciprocity may be regarded as giving each person “reason to believe” that the other has had the same experience. Lewis [14] required reason to believe in his in his definition of “common knowledge”\(^3\). Though we treat shallow beliefs here, we require reason to believe for the formation of beliefs about the other’s beliefs.

Mead [15] has also been influential, in particular, in suggesting the importance of role-switching for obtaining a social perspective. One point of tension and dispute is that thinking about the other’s understanding will lead to cooperation. This idea was emphasized by Mead [15] and his predecessor, Cooley [5], to argue the pervasiveness of cooperation in human society. This was criticized as too naive by later sociologists (see Collins [4], Chap.7). In our theory, cooperation is one possibility obtained by role-switching, but not necessarily guaranteed. We are interested in how role-switching affects the formation of individual preferences and decision making. In particular, we formulate hypotheses about the importance of role-switching for the emergence of cooperation between two persons.

Here, we give a summary of the new concepts that will be used in this paper emphasizing some important results. Since our approach is based on IGT developed in Kaneko-Kline [9], the concepts of a memory kit and inductively derived views (i.d.views) defined there will be taken and adjusted as necessary to fit the current context of role-switching. We remind the reader that IGT starts with the no prior knowledge assumption that each

\(^3\)This is reminiscent of the fixed-point characterization of “common knowledge” (cf., Fagin et al. [6] and Kaneko [8]).
person has little knowledge about the game structure in the beginning. Under this assumption, a person must collect some experiences from playing the game in order to acquire some understanding of the situation.

The new concepts are:

(1): A 2-role Game in a Recurrent Situation and Role-switching: We assume that a particular situation \( G^0 \) in Fig.1.1 is given as a 2-role strategic game \( G^0 \) with the distinction between the roles and persons. Each of the two persons takes one role in each instance of \( G^0 \), and they switch the roles from time to time. Frequency of each role is externally given, but each person has some subjective memory of this frequency. The key departure from the previous work in IGT is the introduction of role-switching.

(2): Transpersonal Understandings: The transpersonal understanding is a new description of a person’s thought of the other’s understanding of the game \( G^0 \). It is added to incorporate his beliefs about the other’s beliefs. This is an extension of the notion of an i.d.view given in Kaneko-Kline [9] in the present context. To avoid confusion, we refer to \( i \)’s own i.d.view as his direct understanding, and his belief about \( j \)’s i.d.view as his transpersonal understanding.

The direct and transpersonal understanding are still static descriptions, though they come from the dynamic interactions in the recurrent situation. When a person adds specific dynamic aspects to the static descriptions, he obtains an inductively derived view, to be defined in Section 4. The dynamic nature of this i.d.view distinguishes it from the i.d.view developed in Kaneko-Kline [9].

Table 1.1: PD

<table>
<thead>
<tr>
<th></th>
<th>( s_{b1} )</th>
<th>( s_{b2} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( s_{a1} )</td>
<td>(5, 5)( \text{ICE} )</td>
<td>(2, 6)( \text{NE} )</td>
</tr>
<tr>
<td>( s_{a2} )</td>
<td>(6, 2)</td>
<td>(3, 3)( \text{NE} )</td>
</tr>
</tbody>
</table>

Table 1.2: SH1

<table>
<thead>
<tr>
<th></th>
<th>( s_{b1} )</th>
<th>( s_{b2} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( s_{a1} )</td>
<td>(7, 7)( \text{NE,ICE} )</td>
<td>(1, 4)</td>
</tr>
<tr>
<td>( s_{a2} )</td>
<td>(4, 1)</td>
<td>(2, 2)( \text{NE} )</td>
</tr>
</tbody>
</table>

Table 1.3: SH2

<table>
<thead>
<tr>
<th></th>
<th>( s_{b1} )</th>
<th>( s_{b2} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( s_{a1} )</td>
<td>(7, 7)( \text{NE,ICE} )</td>
<td>(1, 4)</td>
</tr>
<tr>
<td>( s_{a2} )</td>
<td>(4, 1)</td>
<td>(3, 3)( \text{NE,ICE} )</td>
</tr>
</tbody>
</table>

(3): Intrapersonal coordination equilibrium (ICE): This is defined for the case when the people use both their direct and transpersonal understandings. It coincides with Nash equilibrium in the case of no role-switching. When role-switching is sufficiently reciprocal, it is determined by the unweighted joint payoff sum maximization on the domain over unilateral deviations. For example, in the game of Table 1.1 (Prisoner’s Dilemma), the ICE is given as \( (s_{a1}, s_{b1}) \), which is not an NE. Table 1.2 (Stag Hunt 1) has also \( (s_{a1}, s_{b1}) \) as a unique ICE, which is a NE, while the other NE \( (s_{a2}, s_{b2}) \) is not...
an ICE. However, in Table 1.3 (SH2) which differs only at the payoffs, the strategy pair $(s_{a2}, s_{b2})$ becomes a new ICE. This differences between Nash equilibrium and ICE are fundamental distinctions between IGT with role-switching and standard game theory.

It is important to emphasize that the joint payoff sum maximization needs the use of both the direct and transpersonal understandings. With only the direct understanding, we would again resort to Nash equilibrium behavior (Theorem 4.1).

A key differentiating the ICE from NE is to take a novel effect of role-switching into account; specifically, the fact that each person puts his feet into the other’s shoes affects his decision making as well as his understanding. role-switching also enables each person with an understanding of the social situation including the other’s thoughts. Behaviorally, the ICE leads each person to maximize the weighted payoff sum based on frequencies of roles over the domain of unilateral deviations for both roles.

The precise conditions of reciprocity and experiences for the emergence of this payoff sum maximization are described in Section 5. When those conditions are met, we obtain a utilitarian theorem (Theorem 5.3) which states that the frequency weights of role-switching disappear, and the ICE maximizes the unweighted sum of joint payoffs.

Although we emphasize the case of high reciprocity and the utilitarian theorem, the case of low reciprocity is also considered. In the extreme case of no role-switching, ICE coincides with NE (Theorem 5.1.(2)).

In summary, when reciprocity is low or the transpersonal understanding is ignored (partial use of the i.d.view), we predict a Nash outcome. On the other hand, when role-switching is reciprocal enough, we get the utilitarian result. Thus, we predict the emergence of cooperation from role-switching. This has some implications for the theory of social morality, which are discussed in Section 6. Also, we discuss an implication to the literature of cooperative game theory in Section 8.

Before going to the main body of the paper, we give two comments: one on Lewis’ [14] approach to “common knowledge”, and the other on the repeated game approach (cf., Osborne-Rubinstein [17]), since both seem related.

In our definition of a transpersonal understanding we will require that a person has a reason from experiences to believe the other person has the same beliefs. This requirement is stated by Lewis, though abstractly, to define the concept of common knowledge. In this paper, we stop at shallow depths of interpersonal beliefs, but formulate an experiential version of his idea of a “reason to believe”. A continuation of our research is to study a hierarchy of interpersonal beliefs in the common knowledge logic or universal-type space approach. We discuss this aspect in Section 8.

The repeated game approach is similar to our approach in that both target recurrent situations and discuss cooperation as a possible outcome. Nevertheless, these approaches have radical differences in their basic cognitive postulates. The repeated game approach formulates the entire situation as a huge one-shot game. When cooperation is obtained, it is interpreted as describing ex ante decision making. This requires each player to know
the entire game structure. In our approach, the views of the players are emerging with their experiences. This difference allows us to discuss both the emergence of persons’ understandings of the recurrent situation, as well as the emergence of cooperation.

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 situation and the transpersonal understanding of the other’s understanding, which is an intermediate step to the main definition of an i.d.view given in Section 4. In Section 5, two variants of an intrapersonal coordination equilibrium are defined and studied. In Section 6, we will consider implications of the results obtained in Section 5. In Section 7, we will discuss external and internal reciprocal relations between the persons. In Section 8, we will discuss possible extensions of our approach as well as implications of our cooperation result to some extant game theory literature.

2. Two-Person Strategic Game with Social Roles

In Section 2.1, we give definitions of the 2-role game with role-switching and of a memory kit. In Section 2.2, we discuss the underlying recurrent situation behind a memory kit in an informal manner.

2.1. 2-Role Strategic Game, Role Assignments, and Memory Kits

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, \( S_r = \{s_{r1}, \ldots, s_{r\ell_r}\} \) is a finite set of actions, and \( h_r : S_a \times S_b \to \mathbb{R} \) is a payoff function for role \( r = a, b \). Each role is taken by a person \( i = 1, 2 \). A role assignment \( \pi \) is a one-one mapping from \( \{a, b\} \) to \( \{1, 2\} \), which describes the role taken by each persons in a particular instance of \( G \). The expression \( \pi = (i_a, i_b) \) means that persons \( i_a \) and \( i_b \) take roles \( a \) and \( b \). We use the convention that if \( r = a \) (or \( b \)), then \( s_{(r)} \equiv s_{-r} = s_b \) (or \( s_a \)), but \( (s_r; s_{-r}) = (s_a, s_b) \), and that when we focus on person \( i \), the other person is denoted by \( j \). A 2-person (strategic) game with social roles \( G(\pi) = (i_a, i_b, S_a, S_b, h_a, h_b) \) is given by adding a role assignment \( \pi = (i_a, i_b) \) to a 2-role strategic game \( G \).

Since the situation 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 \( i \) at role \( r \) observes

**Assumption Ob**: the action pair \((s_a, s_b)\) played and his own payoff \( h_r(s_a, s_b) \) in \( G(\pi) \).

Since payoffs represent subjective elements, each person is assumed to observe his own payoff in each play. Here, we assume that each person recognizes each payoff value \( h_r(s_a, s_b) \) only when he experiences it, but he does not know the function \( h_r \) itself.
He may come to know some part of the payoff function only after he has accumulated enough memories.

Now, we consider person $i$’s accumulation of experiences from the 2-role strategic game $G$ up to a particular point of time. They are summarized as a memory kit $k_i = ((s^a_i, s^b_i), (D_{ia}, D_{ib}), (h_{ia}, h_{ib}); (\rho_{ia}, \rho_{ib}))$, which consists of:

$k_1$: the pair $(s^a_i, s^b_i) \in S_a \times S_b$ of regular actions;

$k_2$: the accumulated domain of experiences $D_i = (D_{ia}, D_{ib})$ from taking role $r = a, b$, where $(s^a_{ir}, s^b_{ir}) \in D_{ia} \cup D_{ib} \subseteq S_a \times S_b$;

$k_3$: person $i$’s observed payoff functions $(h_{ia}, h_{ib})$ over $D_i$, where $h_{ir} : D_{ir} \to \mathbb{R}$ and $h_{ir}(s_a, s_b) = h_r(s_a, s_b)$ for all $(s_a, s_b) \in D_{ir}$ and $r = a, b$;

$k_4$: person $i$’s (subjective) frequency weights $(\rho_{ia}, \rho_{ib})$ for roles $a$ and $b$, where $\rho_{ia} + \rho_{ib} = 1$, $\rho_{ia}, \rho_{ib} \geq 0$.

Person $i$ has acquired these by playing game $G$ with different roles from time to time. Condition $k_1$ means that the persons regularly play actions $s^a_i$ and $s^b_i$ when they are assigned to roles $a$ and $b$. Condition $k_2$ states that person $i$ has accumulated experiences $D_{ia}$ and $D_{ib}$ of action pairs from roles $a$ and $b$. We allow $D_{ia}$ or $D_{ib}$ to be empty, though one of them is nonempty since $(s^a_{ir}, s^b_{ir}) \in D_{ia} \cup D_{ib}$. The third components, $(h_{ia}, h_{ib})$, are the perceived payoff functions over $(D_{ia}, D_{ib})$, which are assumed to take the observed values of the payoff functions $(h_a, h_b)$. The last component $(\rho_{ia}, \rho_{ib})$ expresses person $i$’s subjective evaluation of frequencies of roles $a$ and $b$. Accurate weights are not really our intention, but here we assume that it is a single vector for each $i$.

We assume the following on a memory kit:

$$\text{for all } r = a, b, \text{ if } (s_a, s_b) \in D_{ir}, \text{ then } (s_a, s^b_{ir}) \in D_{ir} \text{ and } (s^a_{ir}, s_b) \in D_{ir}; \quad (2.1)$$

$$\rho_{ir} = 0 \text{ if and only if } D_{ir} = \emptyset. \quad (2.2)$$

Condition (2.1) states that if $(s_a, s_b)$ is accumulated in $D_{ir}$, then $(s_a, s^b_{ir})$ and $(s^a_{ir}, s_b)$ coming from the unilateral trials of $s_a$ and $s_b$ from $(s^a_{ir}, s^b_{ir})$ are also accumulated in $D_{ia} \cup D_{ib}$. Condition (2.2) states that $\rho_{ir} = 0$ means that person $i$ has no recollection of being in role $r$. By this, $\rho_{ir} > 0$ if and only if $D_{ir} \neq \emptyset$, which is further equivalent, by (2.1), that $(s^a_{ir}, s^b_{ir}) \in D_{ir}$.

Some cognitive limitations associated with bounded rationality are implicit in our formulation of a memory kit in IGT. In particular, we do not include the entire sequence of past memories in the memory kit. This type of simplification is justified by the cognitive postulates described in Section 2.2 which we attribute to people in IGT.

Using (2.1) twice, we have the following lemma: if person $i$ has some experience at role $r$ in his mind, the pair of regular actions is accumulated at that role.

**Lemma 2.1.** If $D_{ir} \neq \emptyset$, then $(s^a_{ir}, s^b_{ir}) \in D_{ir}$.
When \((s_r; s_{o-r}) \in D_{ir}\), it is called an \textit{active experience} for person \(i\) at role \(r\). That is, if person \(i\) makes a deviation and it remains in his domain, it is an active experience. When \((s_r; s'_{o-r}) \in D_{i(-r)}\), it is a \textit{passive experience} for person \(i\) at role \(-r\).

Reciprocity is important in this paper, but we have various notions and degrees of reciprocity. An important one is between \(D_{ia}\) and \(D_{ib}\) for the same person \(i\). First we define the set \(\text{Proj}(T) := \{(s_a, s_b) : s_a = s'_a \text{ or } s_b = s'_b\}\), where \(T \subseteq S_a \times S_b\) and \((s'_a, s'_b) \in T\). Then, we say that \(D_{ia}\) and \(D_{ib}\) are \textit{internally reciprocal} iff

\[
\text{Proj}(D_{ia}) = \text{Proj}(D_{ib}),
\]

(2.3)

This requires the equivalence of \(D_{ia}\) and \(D_{ib}\) up to unilateral changes from the regular actions \((s'_a, s'_b)\). A stronger reciprocity is \(D_{ia} = D_{ib}\), but (2.3) is more relevant in this paper. The choice of the word “internally” in the definition of (2.3) is to stress that this condition is about reciprocity across domains of a single person \(i\). However, since person \(i\) interacts with person \(j\), this internal reciprocity may be externally motivated. In Section 7, we show how the internalsecondof (2.3) can be derived entirely from external conditions on reciprocity.

By (2.2), \(\rho_r = 0\) is incompatible with (2.3). It would be natural to introduce lower and upper bounds for \(\rho_r\) for these domains. In Section 5.2, we will give bounds when we talk about the utilitarian result (Theorem 5.3).

Consider two examples for the domains \((D_{1a}, D_{1b})\) and \((D_{2a}, D_{2b})\).

(1) \textbf{(Non-reciprocal Domains):} Let \(D_{1}^N = (D_{1a}^N, D_{1b}^N)\) be given as follows:

\[
D_{1a}^N = \{(s_a, s'_b) : s_a \in S_a\} \text{ and } D_{1b}^N = \emptyset.
\]

(2.4)

Let \(D_{2}^N = (D_{2a}^N, D_{2b}^N)\) be defined in a symmetric manner. By (2.2), \(\rho_{1a} = \rho_{2b} = 1\). In this case, each person makes deviations over all his actions, but each accumulates only active experiences: He is either insensitive to (or ignores) the deviations by the other person. Internal reciprocity (2.3) does not hold for these domains.

There are other non-reciprocal domains. For example, each person is sensitive to both active and passive deviations. These domains are not yet internally reciprocal. We have also varieties of reciprocal domains. We focus on one reciprocal case, which satisfies internal reciprocity (2.3).

(2) \textbf{(Active-Passive Domain):} \(D_{1}^{AP} = (D_{1a}^{AP}, D_{1b}^{AP})\) for person 1 is given as:

\[
D_{1a}^{AP} = D_{1b}^{AP} = \{(s_a, s'_b) : s_a \in S_a\} \cup \{(s'_a, s_b) : s_b \in S_b\}.
\]

(2.5)

This domain for player 2 is defined in the same manner. Person 1 makes trials with all actions across both roles, and he is sensitive to 2’s trials as well as his own, but not joint-trials.
2.2. Informal Postulates for Behavior and Accumulation of Memories

Our mathematical theory starts with a memory kit $\kappa_i = \langle (s^a_o, s^b_o), (D_{ia}, D_{ib}), (h_{ia}, h_{ib}); (\rho_{ia}, \rho_{ib}) \rangle$. Behind it, there is some underlying process of behavior and accumulation of memories. We now describe the basic postulates for such a process, which are given in Kaneko-Kline [9] with an addition of BH0 for role-switching.

The additional one is stated explicitly as a postulate.

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

We make the two behavioral postulates: BH1 requires the rule-governed behavior of each person in the recurrent situation ..., $G^o(1,2), G^o(2,1), ..., G^o(1,2)$, ... Postulate BH2 describes how a person makes trials and errors.

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

**Postulate BH2 (Occasional Deviations):** Once in a while (infrequently), each person at 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 $s^r_o$.

In the beginning, each person started behaving almost randomly, and then may have adopted the regular actions $s^a_o$ and $s^b_o$ for roles $a$ and $b$ for some time without thinking, maybe, since he found it worked well in the past or he was taught to follow it. 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.

Learning is possible when there is some regularity; a simple form of regularity is assumed in BH1. Without assuming regular actions and/or patterns, a person may not be able to extract any causality from his experiences. To learn parts other than the regular actions, the persons need to make some trial deviations, which is described by BH2. It may be the case that the regular actions are person-dependent, but as stated in BH1, we restrict our attention to the case where both persons follow the same regular action for each role.

What person $i$ receives in an instant is described by his local (short-term) memory, which 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, his domain $D_{ir}$ 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 $\kappa_i$. For the transition from local to long-term memories, we have various scenarios. Here we list postulates based on bounded memory abilities.

**Postulate EP1 (Forgetfulness):** If experiences are not frequent enough, they would not be transformed into a long-term memory and disappear from a person’s mind.
**Postulate EP2 (Habituation):** A local 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.

When the persons follow their regular actions, the local memories given by them will become long-term memories by EP2. A pair obtained by only one person's deviation from the regular behavior is more likely to remain in his memory than are pairs obtained by joint deviations, which supports (2.1).

A memory kit describes a set of long-term memories, which have been accumulated from the former experiences governed by Postulates EP1 and EP2. This excludes the possibility that a person keeps an entire sequence of his former experiences. To keep a long sequence of experiences needs the person to have a strong memory ability, which is violated by EP1 and EP2. We assume that the long-term memories accumulated from his experiences takes the form of a memory kit as described by $\kappa_1 - \kappa_4^4$.

Those postulates can be tested in experiments; Takeuchi et al. [20] undertook an experimental study of the validity of some of the postulates. We give a brief discussion on this in Section 6.

### 3. Direct and Transpersonal Understandings from Experiences

When a person considers the situation $G$ based on his accumulated experiences, he meets two problems: (i) his own understanding about $G$; and (ii) his thought of the other’s understanding of about $G$. The former is obtained by combining his experiences, while the latter needs some additional interpersonal thinking. In this section, we describe how a person deals with these problems. We do not yet include the regular actions $(s_{ai}, s_{bi})$ and frequency weights $(\rho_{ia}, \rho_{ib})$, which will be taken into account in the definition of an inductively derived view in Section 4. In this sense, the following are static descriptions of the underlying game $G$.

We state our basic ideas for the above mentioned problems as informal postulates before mathematizing them. The first postulate is for the above mentioned (i).

**Postulate DU (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 his own understanding $g^{ii}$ of game $g$. We turn our attention to his thought of the other person’s understanding $g^{ij}$ of $g$. We adopt two postulates for it:

**Postulate TP1 (Projection of Self to the Other):** Person $i$ projects some of his experienced payoffs onto person $j$, when $i$ experientially believes that $j$ has experienced those payoffs.

---

4 An explicit process of transformation from experiences (short-term memories) to long-term memories is given in Akiyama, et al. [1].
By Assumption Ob, he observes only his own payoffs. To think about the other’s payoffs, he uses his experienced payoffs. By postulate TP1, we propose that a person projects his experiences onto the other. Nevertheless, TP1 is a conditional statement. We require some experiential evidence for person $i$ to believe that $j$ knows the payoff, which is expressed as the next postulate.

**Postulate TP2 (Experiential Reason to Believe):** Person $i$ believes that $j$ has experienced payoffs only when $i$ has a sufficient experiential reason for it.

A simple metaphor may help the reader understand those postulates:\footnote{In the example of a baseball team, Mead [15] argued that switching positions often help the players to improve their performance.} A boy notices that a girl appears suffering from the agony of a broken heart. He had experiences of a broken heart a few times, and understands that it is painful. Also, he knows who has caused her broken heart. Here, he projects his former experiences of having pains to her. The ability of projecting his former experiences is stated by TP1, and the reason for her broken heart is required by TP2. In the case where he has no idea of her broken heart, he doubts her behavior; he may think that she is pretending in that manner.

Let us return to our mathematical world. Suppose that person $i$ has accumulated his experiences in a memory kit $\kappa_i = \langle (s_{ia}, s_{ib}), (D_{ia}, D_{ib}), (h_{ia}, h_{ib}); (\rho_{ia}, \rho_{ib}) \rangle$. He now constructs his direct understanding (d-understanding) of the situation from $\kappa_i$ by person $i$ is given as $g_{ii} = (a, b, S_{ia}, S_{ib}, h_{ia}, h_{ib})$:

**Definition 3.1 (Direct and Transpersonal Understandings).** Let a memory kit $\kappa_i = \langle (s_{ia}^0, s_{ib}^0), (D_{ia}, D_{ib}), (h_{ia}, h_{ib}); (\rho_{ia}, \rho_{ib}) \rangle$ be given:

1. **ID1**: The direct understanding (d-understanding) of the situation from $\kappa_i$ by person $i$ is given as $g_{ii} = (a, b, S_{ia}, S_{ib}, h_{ia}^{ii}, h_{ib}^{ii})$:
   - **ID1**: $S_{ia}^i = \{ s_r : (s_r; s_{r-}) \in D_{ia} \cup D_{ib} \text{ for some } s_{r-} \}$ for $r = a, b$;
   - **ID2**: for $r = a, b$, $h_{ir}^{ii} : S_{ia}^i \times S_{ib}^i \to R$ is defined as follows:
     \[
     h_{ir}^{ii}(s_a, s_b) = \begin{cases} 
     h_r(s_a, s_b) & \text{if } (s_a, s_b) \in D_{ir} \\
     \theta_r & \text{otherwise,}
     \end{cases}
     \]
   where the value $\theta_r$ is assumed to satisfy\footnote{This $\theta_r$ can be extended to allow dependence upon persons, and even upon actions, as far as (3.2) is satisfied.}

\[
\theta_r < \min_{(s_a, s_b)} h_r(s_a, s_b). \quad (3.2)
\]
(2): The transpersonal understanding (tp-understanding) from $\kappa_i$ by person $i$ for person $j$ is given as $g^{ij} = (a, b, S^i_a, S^i_b, h^{ij}_a, h^{ij}_b)$, where $h^{ij}_a$ and $h^{ij}_b$ are new and given as:

**ID2** for $r = a, b$, $h^{ij}_r : S^i_a \times S^i_b \to R$ by

$$
h^{ij}_r(s_a, s_b) = \begin{cases} h_r(s_a, s_b) & \text{if } (s_a, s_b) \in D_{ir} \cap D_{i(-r)} \\ \theta_r & \text{otherwise.} \end{cases} \tag{3.3}
$$

All the components of $g^{ij}$ and $g^{ji}$, except $\theta_r$ for the unexperienced part of $S^i_a \times S^i_b$, are determined by memory kit $\kappa_i$. The definition of $g^{ii}$ is straightforward; the $d$-understanding $g^{ii}$ is defined as a 2-role game, based on his experiences. In ID1$^i$, all the experienced actions are taken into account. In ID2$^{ii}$, he constructs his observed payoff function. The value $\theta_r$ expresses an unknown (unexperienced) payoff and is assumed to be small enough which will be important for our existence result of Theorem 5.1.

Consider $g^{ij}$. The difference between $g^{ij}$ and $g^{ji}$ is only in terms the payoffs of (3.1) and (3.3) for an experienced pair $(s_a, s_b)$. For $h^{ij}_r(s_a, s_b) = h_r(s_a, s_b)$, (3.1) require only that $i$ has experienced $(s_a, s_b)$ at the role $r$. However, for $h^{ij}_r(s_a, s_b) = h_r(s_a, s_b)$, (3.3) require that $i$ has experienced $(s_a, s_b)$ at both roles: “$(s_a, s_b) \in D_{ir}$” corresponds to Postulate TP1, i.e., he could project his experience onto the other. But we also require the additional “$(s_a, s_b) \in D_{i(-r)}$” in (3.3), which corresponds to TP2: he should only make this projection if he has the reason to believe that the other has observed his payoff. When $(s_a, s_b) \in D_{ir}$ but $(s_a, s_b) \notin D_{i(-r)}$, person $i$ does not have a reason to believe that $j$ ever experienced payoff $h_r(s_a, s_b)$, and thus, he does not project his payoff experience onto person $j$.

Lewis [14] requires the existence of some reason to believe the other to have the same knowledge for his definition of common knowledge. Postulate TP2 takes the same idea from the experiential perspective. In this paper, we keep the shallow interpersonal beliefs, but not to the common knowledge, since the present framework does not allow us to formulate those explicitly. We will discuss higher order beliefs in Section 8.

Let us exemplify the above definitions with the PD game of Table 1.1 assuming the regular actions $(s^i_0, s^i_2) = (s_{a1}, s_{b1})$:

**3.1 Non-reciprocal Active Domain**: Let $(D^N_{la}, D^N_{lb})$ be given as the non-reciprocal domain of (2.4), where we consider only $G(1, 2)$. Person 1’s $d$-understanding $g^{11} = g^{11}$ is given as: $S^1_a = \{s_{a1}, s_{a2}\}$ and $S^1_b = \{s_{b1}\}$. Since 1 has experiences only for role $a$, the payoffs $(h^{11}_a(s_a, s_b), h^{11}_b(s_a, s_b))$ are given in Table 3.1.

<table>
<thead>
<tr>
<th>Table 3.1; $g^{11}$</th>
<th>Table 3.2; $g^{12}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$s_{a1}$</td>
<td>$s_{b1}$</td>
</tr>
<tr>
<td>$\theta_b$</td>
<td>$\theta_b$</td>
</tr>
<tr>
<td>$s_{a2}$</td>
<td>$s_{b1}$</td>
</tr>
<tr>
<td>$\theta_b$</td>
<td>$\theta_b$</td>
</tr>
</tbody>
</table>
Consider $g^{12}$: Person 1 has experienced the pairs in $D_{1a}^N$, and from each pair, he infers that person 2 observes also these pairs. Hence, 1 can assume the same $S^1_a$ and $S^1_b$ for 2. But person 1 has a difficulty in inferring what payoffs 2 could receive from roles $a$ and $b$. The easier part is $h^{12}_b(s_a, s_b) = \theta_b$ for role $b$ since person 1 has no experiences with role $b$. The other equation $h^{12}_a(s_a, s_b) = \theta_a$ 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 cannot project his experienced payoff onto 2. In sum, person 1 has no idea about person 2’s understanding of payoff values.

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

Lemma 3.1: Let $\rho_{ir} = 1$. Then, $h^{ij}_r(s_a, s_b) = h^{ji}_r(s_a, s_b) = \theta_r$ 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 (2.2). By (3.1) and (3.3), we have the stated equations.  

(2): (Reciprocal Active-Passive Domain): Let $D^A_1 = (D^A_{1a}, D^A_{1b})$ be the domains given by (2.5). Then, $S^i_a = \{s_{a1}, s_{a2}\}$ and $S^i_b = \{s_{b1}, s_{b2}\}$. Both $g^{11}$ and $g^{12}$ are given by Table 3.3. Indeed, person 1 has had each experience along the top row and left column from the perspective of each role. Thus, he can project his experiences onto 2. Only the joint trials are excluded as they are outside the domains of accumulation.

| Table 3.3: $g^{11}$ and $g^{12}$ |
|-----------------|-----------------|
| $a \backslash b$ | $s_{b1}$ | $s_{b2}$ |
| $s_{a1}$ | (5, 5) | (1, 6) |
| $s_{a2}$ | (6, 1) | (\(\theta_a, \theta_b\)) |

Internal reciprocity will be important in our later analysis. We give one theorem that internal reciprocity (2.3) is necessary and sufficient for coincidence of a person’s direct and transpersonal understandings up to the active and passive experiences. Let $g^{ii}$ and $g^{jj}$ be the $d$- and $tp$-understandings from a memory kit $\kappa_i$.

Theorem 3.2 (Internal Coincidence): $(D_{ia}, D_{ib})$ is internally reciprocal if and only if $g^{ii}$ coincides with $g^{jj}$ up to the active/passive experiences, i.e., $h^{ii}_r(s_a, s_b) = h^{jj}_r(s_a, s_b) = h_r(s_a, s_b)$ for all $(s_a, s_b) \in \text{Proj}(S^i_a \times S^i_b)$ and $r = a, b$.

Proof. (Only-if): Suppose $\text{Proj}(D_{ia}) = \text{Proj}(D_{ib})$. Then, we show $\text{Proj}(S^i_a \times S^i_b) = \text{Proj}(D_{ia}) = \text{Proj}(D_{ib})$. Once this is shown, the only-if part is obtained.

Let $(s_a, s_b) \in \text{Proj}(S^i_a \times S^i_b)$. Then, $(s_a \in S^i_a$ and $s_b = s^o_b)$ or $(s_b \in S^i_b$ and $s_a = s^o_a)$. In the first case, $(s_a, t_b) \in D_{ia} \cup D_{ib}$ for some $t_b$. By (2.1), $(s_a, s^o_b) \in D_{ia}$ or $(s_a, s^o_b) \in D_{ib}$. Since $\text{Proj}(D_{ia}) = \text{Proj}(D_{ib})$, we have $(s_a, s^o_b) \in \text{Proj}(D_{ia}) = \text{Proj}(D_{ib})$. In the second case, we have also $(s^o_a, s_b) \in \text{Proj}(D_{ia}) = \text{Proj}(D_{ib})$. Conversely, let $(s_a, s_b) \in \text{Proj}(D_{ia})$.  

13
Then, \((s_a, s_b) \in D_i^r\), which implies \(s_a \in S_a^i\). Similarly, \(s_b \in S_b^i\). Thus, \((s_a, s_b) \in S_a^i \times S_b^i\). Since \(s_a = s_a^o\) or \(s_b = s_b^o\), we have \((s_a, s_b) \in \text{Proj}(S_a^i \times S_b^i)\).

(If): We prove the contrapositive. Suppose \((s_a, s_b) \in \text{Proj}(D_i^r)\) and \((s_a, s_b) \notin \text{Proj}(D_i^{(-r)})\). Since \(h^{ii}_r(s_a, s_b) = h_r(s_a, s_b)\) by (3.1), and \(h^{ij}_r(s_a, s_b) = \theta_r\) by (3.3), we have \(h^{ii}_r(s_a, s_b) = h_r(s_a, s_b) > \theta_r = h^{ij}_r(s_a, s_b)\) by (3.2).

4. Inductively Derived Views for role-switching

The understandings \(g^{ii}\) and \(g^{ij}\) are static descriptions of the recurrent situation. The situation includes temporal aspects such as the regular actions \((s_a^0, s_b^0)\) and frequency weights \((\rho_{ia}, \rho_{ib})\). We define the inductively derived view \(\Gamma^i\) by adding these two components. It contains person \(i\)'s understanding and his belief about \(j\)'s understanding of the recurrent situation. Person \(i\) uses \(\Gamma^i\) to revise his behavior. There are two cases: the partial and full uses of \(\Gamma^i\) for the revision of behavior. In this section, we first present the inductively derived view \(\Gamma^i\), and discuss the partial use of it. The full use is discussed in Section 5.

4.1. Inductively Derived Views

As stated above, the target situation includes temporal aspects in addition to the static descriptions \(g^{ii}\) and \(g^{ij}\). Although making trials and errors is also temporal, it is of a transitory nature. Our concern is to include temporal, but stationary aspects, namely: the regular actions \((s_a^0, s_b^0)\), and the frequency weights \((\rho_{ia}, \rho_{ib})\). We define the weighted payoff functions \(H^{ii}\) and \(H^{ij}\) using \((\rho_{ia}, \rho_{ib})\), where \(H^{ii}\) is the weighted payoff function person \(i\) assigns to himself, and \(H^{ij}\) is the weighted payoff he assigns to \(j\).

To describe these weighted payoffs formally, we introduce the expression \([s_a, s_b]_r\), meaning that person \(i\) at role \(r\) plays \(s_a\) while \(j\) at role \(r\) plays \(s_b\). We consider the recurrent situation where \([s_a, s_b]_a\) is played with frequency \(\rho_{ia}\), and \([t_a, t_b]_b\) is played with frequency \(\rho_{ib} = 1 - \rho_{ia}\). This situation is evaluated by person \(i\) as the weighted\(^7\) payoff:

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

Note that \((s_a, s_b)\) and \((t_a, t_b)\) may be identical.

The same situation is evaluated by \(i\) taking person \(j\)'s perspective as follows:

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

\(^7\)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 [21]. See Hu [7] for a direct approach to expected utility theory from the frequentist perspective.
The difference from (4.1) is that $h^{ij}_a$ and $h^{ij}_b$ are used, and also that weight $\rho_{ia}$ is multiplied to $h^{ij}_b(s_a, s_b)$ and $\rho_{ib}$ to $h^{ij}_a(s_a, s_b)$, since in $i$’s mind, $j$ receives $h^{ij}_b(s_a, s_b)$ with weight $\rho_{ia}$ and $h^{ij}_a(s_a, s_b)$ with weight $\rho_{ib}$.

Although the above definitions look very restrictive relative to the standard definition of the evaluation of an infinitestream of outcomes in the repeated game approach (cf., Osborne-Rubinstein [17]), our definitions of (4.1) and (4.2) are faithful to our motivation to study the emergence of experientially based beliefs obtained by boundedly rational people.

Now, we have the definition of an inductively derived view.

**Definition 4.1.** The inductively derived view (i.d.view) from the memory kit $\kappa_i = \langle (s^a_o, s^b_o), (D_{ia}, D_{ib}), (h_{ia}, h_{ib}); (\rho_{ia}, \rho_{ib}), H^{ii}, H^{ij}\rangle$, where $H^{ii}$ and $H^{ij}$ are the weighted payoff functions given by (4.1) and (4.2) defined\(^8\) over $(S^a_i \times S^b_i)^2$.

The definition of the i.d.view $\Gamma^i$ has various differences from those given in Kaneko-Kline [9], [10] and [11]. One apparent difference is that the above definition is given to a strategic game but not an extensive game (or an information protocol). More importantly, the inclusion of the frequency weights to describe the impact of role-switching is crucial and new.

Though our intention is for person $i$ to use the tp-understanding $H^{ij}$ for his decision making, we should also consider the case where $i$ uses only the direct understanding for his behavior. We divide our analysis into two cases:

**C0(Partial Use):** Person $i$ uses only the payoff function $H^{ii}$.

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

Either C0 or C1 may be taken as a decision criterion for person $i$. The first case of C0 is handled in the next subsection where we find that it leads to Nash equilibrium type behavior. The second case of C1 is more novel and is handled in Section 5. There we see the full force of role-switching as an experiential source for a player’s beliefs about the other’s beliefs.

**4.2. Partial Use of the I.D.View**

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 $H^{ii}$, we need some behavioral assumption about the other person’s action or reaction to his change. We adopt the following conjectural postulate by person $i$ when he takes an intensional deviation:

\(^8\)It is our intention to define $H^{ii}$ and $H^{ij}$ over the unilateral domain $\{(s^a_o; s^o_{-r}) : (s^r_o; s^o_{-r}) \in D_{ir}, r = a, b\}$. However, we define them over $(S^a_i \times S^b_i)^2$ to avoid notational complications.
or if and only if it is a Nash equilibrium in the base game.

We require the present regular actions \((s^o_a, s^o_b)\) are free from such behavior revisions. Thus, we have the following definition: \((s^o_a, s^o_b)\) is a partial-use equilibrium (PUE) in \(\Gamma^i\) iff for all \(s_a \in S^i_a\) and \(s_b \in S^i_b\),

\[
H^{ii}([s^o_a, s^o_b]_a, [s^o_a, s^o_b]_b) \geq H^{ii}([s^o_a, s^o_b]_a, [s^o_a, s^o_b]_b).
\]

That is, person \(i\) maximizes his direct understanding \(H^{ii}\), by controlling \(s_a\) or \(s_b\) when he takes role \(a\) or \(b\), respectively.

We have the following theorem:

**Theorem 4.1 (Partial-Use Equilibrium).** The regular pair \((s^o_a, s^o_b)\) is a PUE in \(\Gamma^i\) if and only if it is a Nash equilibrium in the d-understanding \(g^{ii}\).

**Proof.** The definition of a PUE is expressed as: for all \(s_a \in S^i_a\) and \(s_b \in S^i_b\),

\[
\rho_{ia} h^{ii}_a(s^o_a, s^o_b) + \rho_{ib} h^{ii}_b(s^o_a, s^o_b) \geq \rho_{ia} h^{ii}_a(s_a, s^o_b) + \rho_{ib} h^{ii}_b(s^o_a, s_b).
\]

By this, the if part is straightforward. We show the contrapositive of the only-if part. Suppose that \((s^o_a, s^o_b)\) is not a Nash equilibrium in \(g^{ii}\). Then, there is some \(s_a \in S^i_a\) or \(s_b \in S^i_b\) such that \(h^{ii}_a(s^o_a, s^o_b) < h^{ii}_a(s_a, s^o_b)\) or \(h^{ii}_b(s^o_a, s^o_b) < h^{ii}_b(s^o_a, s_b)\), respectively. If \(h^{ii}_a(s^o_a, s^o_b) < h^{ii}_a(s_a, s^o_b)\), then \(\rho_{ia} > 0\) by (2.2) and (3.1), and similarly if \(h^{ii}_b(s^o_a, s^o_b) < h^{ii}_b(s^o_a, s_b)\), then \(\rho_{ib} > 0\). In the former case, (4.4) does not hold if we plug \((s_a, s^o_b)\) to the first term of the right-hand side but \((s^o_a, s^o_b)\) to the second term. The latter case is parallel. \(\blacksquare\)

This states that the regular actions are free from behavior revisions based on the d-understanding \(g^{ii}\) if and only if the regular action pair \((s^o_a, s^o_b)\) is a Nash equilibrium in \(g^{ii}\). Theorem 4.1 holds with no additional restriction on the accumulated domain of experiences, or on the frequency weights. Our intention for CO is mainly to handle the non reciprocal cases where \(\rho_{ir} = 1\) for some \(r = a, b\). When, this happens, it would be natural to assume that \(\rho_{jr} = 0\). In this case, we can apply the PUE concept to both persons.

To consider this application, first, we assume that the i.d.views of the two persons are coherent with respect to frequency weights. We say that \(\Gamma^i = ((s^o_a, s^o_b), (S^i_a, S^i_b), (\rho_{ia}, \rho_{ib}), H^{ii}, H^{ij}), i = 1, 2\) are mutually coherent iff \(\rho_{ia} = 1 - \rho_{ja}\). Also, we say that the pair \((s^o_a, s^o_b)\) of regular actions is a mutual PUE iff it is a PUE for \(\Gamma^i\) and \(i = 1, 2\). Then we have the following corollary from Theorem 5.1.

**Corollary 4.2 (Mutual PUE for Non-reciprocal Cases).** Let \(\Gamma^1, \Gamma^2\) be the mutually coherent i.d.views with \(\rho_{1a} = 1, S^1_a = S_a\) and \(S^2_b = S_b\). Then, \((s^o_a, s^o_b)\) is a mutual PUE if and only if it is a Nash equilibrium in the base game \(G\).
Proof. Since $\rho_{1a} = \rho_{2b} = 1$ and $S_a^1 = S_a$ and $S_b^2 = S_b$, it holds that $(s_a, s_b^0) \in D_{1a}$ for all $s_a \in S_a$ and $(s_a^0, s_b) \in D_{2b}$ for all $s_b \in S_b$. These imply $h_{a1}^{11}(s_a, s_b^0) = h_a(s_a, s_b^0)$ for all $s_a \in S_a$ and $h_{b2}^{22}(s_a^0, s_b) = h_b(s_a^0, s_b)$ for all $s_b \in S_b$. These together with Theorem 4.1 imply that $(s_a^0, s_b^0)$ is a Nash equilibrium in the base game $G$. ■

5. Intrapersonal Coordination Equilibria

Our main concern is the full use of the i.d. view $\Gamma^i = ((s_a^0, s_b^0), (S_a^i, S_b^i), (\rho_{ia}, \rho_{ib}), H^{ii}, H^{ij})$ for person $i$. Since person $i$ can use both $H^{ii}$ and $H^{ij}$, we should reconsider the conjectural postulate $(\ast)$ for his behavior revision. In Section 5.1, we suggest an alternative conjectural postulate. This leads to an equilibrium concept called $ICE$. In the remaining, we study this equilibrium concept, and obtain some utilitarian result in reciprocal domains.

5.1. Intrapersonal Coordination Equilibria

It is a salient difference between the partial use $C_0$ and the full use $C_1$ of the i.d.view $\Gamma^i$ that person $i$ does not think about person $j$’s perspective in the former, but he does in the latter. In $C_0$, $(\ast)$ would be a natural conjectural postulate. If we stick to $(\ast)$, we would effectively return to the Nash equilibrium (Theorem 4.1) even for the case of reciprocal role-switching. In the full use $C_1$, however, person $i$ can think about how person $j$ thinks about and responds to person $i$’s possible deviations. This leads him to a different conjectural postulate.

It is an alternative postulate that if person $i$ deviates from $s_{o_i}$ to some $s_r$, then

$(\ast\ast)$: person $j$ also takes the same action $s_r$ when $j$ takes role $r$.

This may be called the role model postulate. Person $i$ needs a justification for why he expects person $j$ to follow $i$’s deviation. He may find a justification for it since now he can think about person $j$’s perspective by his i.d. view $\Gamma^i$. We still need two steps to explain this justification.

First, we explain it in a static manner, which takes place in person $i$’s mind. This explains how both persons could be better off following $(\ast\ast)$. But since person $j$, even in the mind of person $i$, should be independent from person $i$, $(\ast\ast)$ requires some external coordination. For this, we refer to the role model argument: One person’s deviation shows an initiative to the other person to deviate in a coordinate way, and expects the other person understand and follow it. This is the second step, for which the restriction on unilateral deviations, excluding joint deviations, is needed.

Suppose that both $i$ and $j$ would get higher payoffs if each deviates from $s_{o_i}^0$ to $s_a$ when each is assigned to role $a$. Person $i$ understands this based on his i.d. view $\Gamma^i$, and also, person $i$ may understand that person $j$ reaches the same understanding. Then
person \( i \) justifies postulate (**) . This argument is formally described in the language of \( \Gamma^i \) : The first part of person \( i \)'s understanding is expressed as:

\[
H^{ii}( [s^o_a, s^0_b]_a, [s^o_a, s^0_b]_b ) < H^{ii}( [s_a, s^0_b]_a, [s_a, s^0_b]_b ). \tag{5.1}
\]

The assumption that \( j \) also switches at role \( a \) is expressed by \([s_a, s^0_b]_b\) on the right hand side of the inequality. The second part of person \( i \)'s belief on \( j \)'s understanding is expressed as:

\[
H^{ij}( [s^o_a, s^0_b]_a, [s^o_a, s^0_b]_b ) < H^{ij}( [s_a, s^0_b]_a, [s_a, s^0_b]_b ). \tag{5.2}
\]

When both (5.1) and (5.2) are satisfied for a deviation from \( s^o_a \) to \( s_a \), person \( i \) justifies postulate (**) .

We say that \( s_r \in S^i_r \) is a coordinately improving deviation (c-improving deviation) from the regular actions \((s^o_a, s^0_b)\) iff (5.1) and (5.2) hold with the replacement of \( s_a \) by \( s_r \). We allow the weaker form of (5.1) and (5.2): That is, we consider a deviation in the form that both (5.1) and (5.2) hold in weak inequalities with a strict inequality for at least one person. When these hold, we call \( s_r \in S^i_r \) a weakly c-improving deviation. Allowing this weak c-improving deviations, we avoid some difficulties arising when \( H^{ii} \) or \( H^{ij} \) takes constant values.

The above argument itself does not exclude the possibility of joint deviations by the two persons. However, such deviations require higher-order expectations. We avoid this problem by the restriction on unilateral deviations, which is also related to the second step of the justification of postulate (**) . Postponing this argument later, we now state the definition of our equilibrium concept in the full use of an i.d.view \( \Gamma^i = ((s^o_a, s^0_b), (S^i_a, S^i_b), (\rho_{ia}, \rho_{ib}), H^{ii}, H^{ij}) \):

**Definition 5.1 (ICE).** The regular pair \((s^o_a, s^0_b)\) is an intrapersonal coordination equilibrium (ICE) in \( \Gamma^i \) iff there is no weak c-improving unilateral deviation from \((s^o_a, s^0_b)\).

This concept requires the regular pair \((s^o_a, s^0_b)\) to be stable against such deviations. The term “intrapersonal coordination” is motivated by the fact that possible deviations are all considered by person \( i \)'s mind. The coordination issue is also related to this term, and is based on the role model argument. Now, let us consider it.

\[
\begin{array}{c|c|c}
1 & 2 \\
\hline
s_a & s^0_b & s^0_b \\
\end{array}
\]

\[
\begin{array}{c|c|c}
1 & 2 \\
\hline
s^0_a & s^0_b & s^0_b \\
\end{array}
\]

\[
\begin{array}{c|c|c}
2 & 1 \\
\hline
s_a & s^0_b & s^0_b \\
\end{array}
\]

\[
\begin{array}{c|c|c}
2 & 1 \\
\hline
s^0_a & s^0_b & s^0_b \\
\end{array}
\]

Fig.5.1

Suppose that person 1 at role \( a \) deviates from \( s^o_a \) to \( s_a \) based on his evaluations of the conjectured improving deviation expressed in terms of (5.1) and (5.2). Since person 1 can change his action at one role at one time, this new situation is expressed as the
second left state in Fig.5.1. Then person 2 will observe this deviation, and when he is assigned to role a, he could follow this mutually beneficial deviation $s_a$, which is the third state in Fig.5.1. Thus, person 1 can take an initiative to deviate from $(s_a^o, s_b^o)$ to $s_a^!$: This is exactly the role model argument. We should remark that the argument is about the external world, but, is thought by person $i$ using his view $\Gamma^i$.

In the end of Section 5.1, we will give comments on a possible alternation of the above definition to allow joint deviations at both roles simultaneously to $(s_a, s_b)$ from $(s_a^o, s_b^o)$, and on coherency with the definition of a PUE.

Our first formal result is about the existence of an ICE. The first assertion is the main point of Section 5.1, while the second is self-explanatory.

**Theorem 5.1 (Existence).** Let $\Gamma^i$ be the i.d.view from a memory kit $\kappa_i = ((s_a^o, s_b^o), (D_{ia}, D_{ib}), (h_{ia}, h_{ib}), (\rho_{ia}, \rho_{ib})$. Let $\rho_{ia} < 1$ and $(s_{ia}^o, s_{ib}^o)$ a pair satisfying

$$h_a(s_{ia}^o, s_{ib}^o) + h_b(s_{ia}^o, s_{ib}^o) \geq h_a(s_r; s_{-r}^*) + h_b(s_r; s_{-r}^*)$$

for all $s_r \in S_r$ and $r = a, b$. (5.3)

Then, $(s_{ia}^o, s_{ib}^o)$ is an ICE in the i.d.view $\Gamma^i$. Thus, $(s_{ia}^o, s_{ib}^o)$ is an ICE in the i.d.view $\Gamma^i$ if and only if it is a Nash equilibrium in person $i$'s d-understanding $g^i$.

**Proof.** (1): Since $0 < \rho_{ia} < 1$, we have $D_{ia} \neq \emptyset$ and $D_{ib} \neq \emptyset$ by (2.2). By (2.1), this implies $(s_{ia}^o, s_{ib}^o) \in D_{ia} \cap D_{ib}$.

Suppose that $(s_{ia}^o, s_{ib}^o)$ is not an ICE. Then, there is some weak $c$-improving unilateral deviation from $(s_{ia}^o, s_{ib}^o)$. Let $s_a$ be such a deviation. The other cases are symmetric. In the present case we have:

$$H^{i\mid}(s_{ia}^o, s_{ib}^o) \leq H^{i\mid}(s_{ia}^o, s_{ib}^o);$$

$$H^{i\mid}(s_{ia}^o, s_{ib}^o) \leq H^{i\mid}(s_{ia}^o, s_{ib}^o),$$

where at least one holds with a strict inequality. Since $(s_{ia}^o, s_{ib}^o) \in D_{ia} \cap D_{ib}$, we have

$$h^i_r(s_{ia}^o, s_{ib}^o) = h_r(s_{ia}^o, s_{ib}^o)$$

and $h^i_r(s_{ia}^o, s_{ib}^o) = h_r(s_{ia}^o, s_{ib}^o)$ for $r = a, b$. If $(s_{ia}^o, s_{ib}^o) \notin D_{ia} \cap D_{ib}$, then $h^i_r(s_{ia}, s_{ib}) = \theta_r$ for $r = a, b$, so the second inequality in (5.4) becomes

$$(1 - \rho_{ia})h_a(s_{ia}^o, s_{ib}^o) + \rho_{ia}h_b(s_{ia}^o, s_{ib}^o) \leq (1 - \rho_{ia})h_a(s_{ia}, s_{ib}) + \rho_{ia}h_b(s_{ia}, s_{ib}),$$

which is impossible by (3.2). Hence, $(s_{ia}^o, s_{ib}^o) \in D_{ia} \cap D_{ib}$. Thus, the two inequalities of (5.4) are expressed as:

$$\rho_{ia}h_a(s_{ia}^o, s_{ib}^o) + (1 - \rho_{ia})h_b(s_{ia}^o, s_{ib}^o) \leq \rho_{ia}h_a(s_{ia}, s_{ib}^o) + (1 - \rho_{ia})h_b(s_{ia}, s_{ib}^o)$$

and

$$(1 - \rho_{ia})h_a(s_{ia}^o, s_{ib}^o) + \rho_{ia}h_b(s_{ia}^o, s_{ib}^o) \leq (1 - \rho_{ia})h_a(s_{ia}, s_{ib}^o) + \rho_{ia}h_b(s_{ia}, s_{ib}^o),$$

where at least one holds with a strict inequality. Summing up these inequalities, we have

$$h_a(s_{ia}^o, s_{ib}^o) + h_b(s_{ia}^o, s_{ib}^o) < h_a(s_{ia}, s_{ib}^o) + h_b(s_{ia}, s_{ib}^o),$$

which is a contradiction to the choice of
(\(s^o_a, s^o_b\)) in (5.3).

(2): Since \(\rho_{ir} = 1\), we have \(D_{i(-r)} = \emptyset\) by (2.2). By this, we observe \(h^{ij}_{ir}(s_r; s_{-r}) = \theta_r\) for all \((s_r; s_{-r}) \in S^i_a \times S^i_b\). That is, \(h^{ij}_{ir}\) is constant, which implies that \(H^{ij}\) is constant. Let \((s^o_a, s^o_b)\) be an ICE in \(\Gamma^i\). Since \(H^{ij}\) is constant, \((s^o_a, s^o_b)\) being an ICE implies that \(H^{ij}\) is maximized at \((s^o_a, s^o_b)\), which implies \(h^{ij}_{ir}(s_r; s^o_{-r}) \leq h^{ij}_{ir}(s^o_r; s^o_{-r})\) for all \(s_r \in S^i_r\). The converse is obtained by tracing this argument back.

By Theorem 5.1.(1), we have the existence result for the case of \(0 < \rho_{ia} < 1\). An algorithm to find a pair \((s^*_a, s^*_b)\) satisfying (5.3) is constructed as follows: Take any pair in the payoff matrix. Then, if there is one pair with a higher sum of payoffs by one person’s deviation, we move to this pair as a candidate for \((s^*_a, s^*_b)\). If this pair has still 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 maximum, but it will satisfy (5.3); thus we have a general existence result for this case.

Theorem 5.1.(1) allows arbitrary \((D_{ia}, D_{ib})\), in which case \(S^i_a\) and \(S^i_b\) may be proper subsets of \(S_a\) and \(S_b\). If this is the case, we can relax (5.3) so that it is a maximization condition relative to proper subsets of \(S_a\) and \(S_b\). This generalization can be done without much difficulty. However, we should notice that we have the trivial extreme case where \(D_{ia} = D_{ib} = \{(s^*, s^*)\}\). Apparently, this is not our target. In Section 5.2, we will restrict our target to the case where \(S^i_a = S_a\) and \(S^i_b = S_b\).

For the non-reciprocal case of \(\rho_{ir} = 1\) for \(r = a\) or \(b\), the ICE is equivalent to NE in his own understanding \(g^{ii}\) by Theorem 5.1.(2). Since \(g^{ii}\) is effectively one-person game in the sense that \(h^{ii}_{ir}\) is constant, the NE in \(g^{ii}\) is simply determined by the payoff maximization of \(h^{ii}_{ir}\), and person \(j\)’s behavior is unconstrained. In the PD game of Table 1.1 with \(S^i_r = \{s_r1, s_r2\}\), an ICE is either \((s_r1, s_{(-r)}1)\) and \((s_r2, s_{(-r)}2)\).

Finally, let us return to the problems of joint deviations and of coherency between a PUE and an ICE. In the restriction of an PUE, a person is allowed to change his action at each role, provided that the other sticks to the regular action \(s^a_r\) at role \(r = a, b\). Without external communication, person 1 can deviates alone, and this deviation induces the second state \((s^o_a, s^o_b)\) of Fig.5.1. In the argument for ICE, person 1 is assumed to return to \(s^o_b\) in the third state \((s^o_a, s^o_b)\) when he is assigned to role \(b\). A natural question is why we avoid the assumption like a PUE that 1 takes a new deviation \(s_b\) in the third state, which is also the main problem of a joint deviation.

Let us recall that the entire argument is considered in the mind of person 1. Thus, he expects that person 2 understands his deviation in the second state as a sign to initiate a coordinate deviation \(s_a\) from \((s^o_a, s^o_b)\). However, if he deviates again from \(s^o_b\) to \(s_b\) in the third state, the previous sign may be destroyed. Logically, it is possible to expect that person 2 understands such consecutive signs, but this requires higher-order expectations. We avoid the argument of this consecutive signs as difficult. On the other
hand, the definition (4.3) of a PUE does not involve this difficulty, since the other’s understanding is not involved at all. However, when the underlying game is given as an extensive game, some external communication may be possible, which will be discussed briefly for the example of an ultimatum game in Section 6.

We remark that a joint deviation argument may have some validity in a small game such as a $2 \times 2$ game. As will be discussed, the Stag-Hunt game 2 of Table 1.3 has two possible ICE’s, but if we allow joint deviations, only one would remain.

In a parallel manner to a mutual PUE, we have a mutual ICE: that the pair $(s^o_a, s^o_b)$ of regular actions is a mutual ICE iff it is an ICE for $\Gamma^i$ and $i = 1, 2$. We have the following corollary from Theorem 5.1.

**Corollary 5.2 (Existence of a Mutual ICE):** Let $\Gamma^i$ be the mutual coherent i.d. views for $i = 1, 2$.

1. Let $0 < \rho_{ia} < 1$ for $i = 1, 2,$ and suppose that $(s^o_a, s^o_b) = (s^*_a, s^*_b)$ satisfies (5.3). Then, $(s^o_a, s^o_b)$ is a mutual ICE.

2. Let $\rho_{ia} = \rho_{ib} = 1$. Assume (5.5) for each $\Gamma^i$. Then, $(s^o_a, s^o_b)$ is a mutual ICE if and only if it is a Nash equilibrium in the game $G$.

### 5.2. Utilitarian Theorem

A pair $(s^*_a, s^*_b)$ given by (5.3) is always an ICE of $\Gamma^i = \langle (s^o_a, s^o_b), (S^i_a, S^i_b), (\rho_{ia}, \rho_{ib}), H^{ii}, H^{ij} \rangle$ with $(s^o_a, s^o_b) = (s^*_a, s^*_b)$ for any frequency weight $\rho_{ia} \in (0, 1)$. Since such a pair has a special status in our theory, we call it a unilateral utilitarian point (UUP). Specifically, for cases of $\rho_{ia}$ near $1/2$, we expect the converse of Theorem 5.1.(1), i.e., every ICE is a UUP, though rigorously speaking, we need some assumptions for it.

First of all, we restrict our attention to the internally reciprocal domains (2.3). Still, the domains $D_{ia}$ and $D_{ib}$ are too arbitrary, as pointed out after Theorem 5.1. To simplify our argument, we assume

\[ S^i_a = S_a \text{ and } S^i_b = S_b. \]  

(5.5)

We write another condition to be used in the next theorem: for all distinct $(s_a, s_b), (s'_a, s'_b) \in S_1 \times S_2$ with $s_r = s'_r$ for $r = a$ or $b$,

\[ h_a(s_a, s_b) + h_b(s_a, s_b) \neq h_a(s'_a, s'_b) + h_b(s'_a, s'_b). \]  

(5.6)

That is, the payoff sum differs for different pairs of unilaterally different actions.

When $\Gamma^i$ is given as $\langle (s^o_a, s^o_b), (S^i_a, S^i_b), (\rho_{ia}, \rho_{ib}), H^{ii}, H^{ij} \rangle$, we denote, by $\Gamma^i(\hat{\rho}_{ia}, \hat{\rho}_{ib})$, the i.d. view obtained from $\Gamma^i$ with the replacement of $(\rho_{ia}, \rho_{ib})$ by $(\hat{\rho}_{ia}, \hat{\rho}_{ib})$.

**Theorem 5.3 (Utilitarian Theorem).** Let $\Gamma^i$ be the i.d. view from a memory kit $\kappa_i$ satisfying (2.3) and (5.5).
(1): \((s'_a, s'_b)\) is an ICE in \(\Gamma^i(\frac{1}{2}, \frac{1}{2})\) if and only if it is a UUP.

(2): Suppose (5.6). Then, \((s'_a, s'_b)\) is a UUP if and only if there are \(\alpha, \beta\) with \(0 < \alpha < 1/2 < \beta < 1\) such that \((s'_a, s'_b)\) is an ICE of \(\Gamma^i(\hat{\rho}_{ia}, \hat{\rho}_{ib})\) for any \(\hat{\rho}_{ia} \in [\alpha, \beta]\).

Proof (1): The if part was already obtained by the proof of Theorem 5.1.(1). Suppose that \((s'_a, s'_b)\) is an ICE in \(\Gamma^i(\frac{1}{2}, \frac{1}{2})\). Since \(\text{Proj}(D_{ia}) = \text{Proj}(D_{ib})\) by (2.3), we have \((s_r; s'_{r}) \in D_{ia} \cap D_{ib}\) for all \(s_r \in S_r, r = a, b\). Hence \(h^i_{\alpha}\) and \(h^i_{\beta}\) coincide with \(h_r\) for the domain of unilateral deviations. Thus, the condition of (5.7) is written as \(\frac{1}{2}h_a(s'_a, s'_b) + \frac{1}{2}h_b(s'_a, s'_b) \geq \frac{1}{2}h_a(s_r; s'_{r}) + \frac{1}{2}h_b(s_r; s'_{r})\) for all \(s_r \in S_r, r = a, b\).

This means that \((s'_a, s'_b)\) is a UUP.

(2): Let \((s'_a, s'_b)\) be a UUP. Then, by (5.6), we have \(\frac{1}{2}h_a(s'_a, s'_b) + \frac{1}{2}h_b(s'_a, s'_b) > \frac{1}{2}h_a(s_r; s'_{r}) + \frac{1}{2}h_b(s_r; s'_{r})\) for all \(s_r \in S_r - \{s'_{r}\}, r = a, b\). Hence, there are some \(\alpha, \beta\) with \(0 < \alpha < 1/2 < \beta < 1\) such that for any \(\hat{\rho}_{ia} \in [\alpha, \beta]\),

\[
\hat{\rho}_{ia}h_a(s'_a, s'_b) + (1 - \hat{\rho}_{ia})h_b(s'_a, s'_b) \geq \hat{\rho}_{ia}h_a(s_r; s'_{r}) + (1 - \hat{\rho}_{ia})h_b(s_r; s'_{r});
\]

\[
(1 - \hat{\rho}_{ia})h_a(s'_a, s'_b) + \hat{\rho}_{ia}h_b(s'_a, s'_b) \geq (1 - \hat{\rho}_{ia})h_a(s_r; s'_{r}) + \hat{\rho}_{ia}h_b(s_r; s'_{r}).
\]

for all \(s_r \in S_r, r = a, b\), and at least one is a strict inequality. These mean that \((s'_a, s'_b)\) is an ICE of \(\Gamma^i(\hat{\rho}_{ia}, \hat{\rho}_{ib})\) for any \(\hat{\rho}_{ia} \in [\alpha, \beta]\).

The converse can be obtained by tracing back this argument. That is, by summing up the inequalities of (5.7), we have the condition for \((s'_a, s'_b)\) to be an UUP.

The counterpart of Corollary 5.2 for Theorem 5.3 can be stated in a similar manner.

The point of Theorem 5.3 is the equivalence between the ICE and UUP under the reciprocal domain (2.3) for \(\hat{\rho}_{ia} = \frac{1}{2}\) or in some neighborhood of \(\frac{1}{2}\). Since (5.6) is naturally expected, we would have the equivalence even for skewed weights. However, the UUP is determined by maximization of the simple sum of payoffs independent of skewed weights: Although person \(i\) detects some skewness of frequencies, the resulting outcome is free from it. A simple utilitarian (up to unilateral domains) outcome is an ICE, which motivates the title of the Theorem 5.3.

It may be questioned how large the size of the interval \([\alpha, \beta]\) in Theorem 5.3.(2) is, and also what would happen outside the interval. We consider these problems in the PD game and the SH games. For this consideration, let us denote the infimum and supremum of such \(\alpha\’s\) and \(\beta\’s\) in Theorem 5.3.(2) by \(\alpha_0\) and \(\beta_0\), respectively. These can be calculated from the inequality system (5.7).

It would be convenient to introduce one definition: We say that \((s'_a, s'_b)\) is an ICE point for \((\hat{\rho}_{ia}, \hat{\rho}_{ib})\) if and only if \((s'_a, s'_b)\) is an ICE in \(\Gamma^i(\hat{\rho}_{ia}, \hat{\rho}_{ib})\) with \(D_{ia} = D_{ib} = \{(s_r; s'_{r}) : s_r \in S_r, r = a, b\}\). Since the boundary case of \(\hat{\rho}_{ir} = 1\) was already discussed in Section 5.1, we consider only the case where \(0 < \hat{\rho}_{ia} < 1\).

Prisoner’s Dilemma: Consider Table 1.1. By Theorem 5.1.(1), the regular pair
(s_{a1}, s_{b1}) is an ICE point for any \( (\hat{\rho}_{ia}, \hat{\rho}_{ib}) \). Since \( (s_{a1}, s_{b1}) \) is the unique UUP, it follows by Theorem 5.3.(2) that there is some interval \([\alpha, \beta]\) such that for \( \hat{\rho}_{ia} \in [\alpha, \beta] \), \( (s_{a1}, s_{b1}) \) is the unique ICE of \( \Gamma'(\hat{\rho}_{ia}, \hat{\rho}_{ib}) \). This holds up to \([\alpha_0, \beta_0] = [\frac{1}{4}, \frac{3}{4}]\).

For any \( \hat{\rho}_{ia} \in (0, \frac{1}{4}) \cup (\frac{3}{4}, 1) \), the other three pairs \( (s_{a1}, s_{b2}) \), \( (s_{a2}, s_{b2}) \) and \( (s_{a1}, s_{b2}) \) appear as ICE points. Thus, all the pairs are ICE's in the skewed weight case.

**Stag Hunt**: The SH1 game of Table 1.2 has the unique UUP \( (s_{a1}, s_{b1}) \). This is the unique ICE for \( \hat{\rho}_{ia} \in [\frac{1}{3}, \frac{2}{3}] \). For any \( \hat{\rho}_{ia} \in (0, \frac{1}{3}) \cup (\frac{2}{3}, 1) \), the other \( (s_{a2}, s_{b2}) \) appears as an ICE point, which is also a NE.

Consider the SH2 game of Table 1.3. Here we have two UUP's, \( (s_{a1}, s_{b1}) \) and \( (s_{a2}, s_{b2}) \). Thus, both are ICE points for any frequency \( \hat{\rho}_{ia} \in (0, 1) \). We have no other candidates for an ICE point. As stated in the end of Section 5.1, if we allow joint deviations, we expect only \( (s_{a1}, s_{b1}) \) for \( \hat{\rho}_{ia} \) near \( \frac{1}{2} \).

Observe the discontinuity of ICE’s at \( \hat{\rho}_{ia} = 0 \) or 1. Condition (2.2) is only a restriction on the relationship between the frequency weight \( \hat{\rho}_{ia} \) and domain \( D_{ir} \); hence, we can assume \( \hat{\rho}_{ia} \in (0, 1) \) and the unilateral domain condition (2.3). However, \( \hat{\rho}_{ia} \) is a subjective evaluation of the objective frequency, and a small and precise value for \( \hat{\rho}_{ia} \) is not very compatible with the bounded cognitive ability of a person. Hence, the objective frequency of role \( a \) near 0 or 1 may be effectively understood as \( \hat{\rho}_{ia} = 0 \) or 1. Then, Theorem 5.1.(2) should be applied to such cases, rather than Theorem 5.3.

To analyze this problem, a simulation study like Akiyama, et al. [1] or an experimental study like Takeuchi et al. [20] could give some information.

### 6. Extensions and Further Applications

We have already seen some applications of our theory to the PD and SH games in Section 4. In this section, first we mention some experimental results on PD games with role-switching. Next, we apply our theory to an Ultimatum Game. Finally, we discuss implications of our theory to moral philosophy.

**Experimental Study**: Takeuchi et al. [20] undertook experiments for the cases of no role-switching and full role-switching for some PD games. They address the question about subjects' behaviors and cognitive understandings of payoff values.

In the case of no role-switching, the experimental results are quite consistent with our theory effectively suggesting the Nash equilibrium. In the case of role-switching, the ICE and NE in addition to nonconvergent behaviors are observed, which are quite consistent with the present theory of role-switching. A salient point of an experiment is to enable us to study the behavioral and cognitive postulates, in particular, how the process converges from the phase of trial and error to equilibrium, and how persons learn the payoffs. From the answers to the questionnaire given after the experiment, we analyzed the relationship between the payoff understandings and their behaviors. In many cases,
Postulates BP1, BP2 and Postulates EP1, EP2 are supported and sometimes sharpened. Those findings complement our theoretical study.

**Ultimatum Game**: A person assigned to role $a$ proposes a division $(x_a, x_b)$ of $\$100$ to persons 1 and 2, and a person assigned to $b$ 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 if the person at $b$ chooses $(\alpha_1, \alpha_2, \alpha_3)$, the outcome depends only upon $\alpha_1$; if $\alpha_1 = Y$, they receive $(99, 1)$ and if $\alpha_1 = N$, they receive $(0, 0)$. For the other cases, we define payoffs in a parallel manner. The game is depicted in Fig.6.1.

This game has a unique backward induction solution: $((99, 1), (Y, Y, Y))$. This is quite incompatible with experimental results (cf., Camerer [3]), which have indicated that $(50, 50)$ is more likely chosen by the person at $a$.

Here, 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 ICE.

Under the assumption that person $i$ has the reciprocal active-passive domains $D_{ia}^{AP} = D_{ib}^{AP}$ and $\rho_{ia} = 1/2$, the pair $((99, 1), (Y, Y, Y))$ is not an ICE, since

\[
\frac{1}{2}h_a((99, 1), (Y, Y, Y)) + \frac{1}{2}h_b((99, 1), (Y, Y, Y)) = \frac{1}{2}u(99) + \frac{1}{2}u(1) < u(50) = \frac{1}{2}u(50) + \frac{1}{2}u(50) = \frac{1}{2}h_a((50, 50), (Y, Y, Y)) + \frac{1}{2}h_b((50, 50), (Y, Y, Y)).
\]

The inequality follows the strict concavity of $u$. In this game, an ICE is given as $((50, 50), (\alpha_1, Y, \alpha_3))$, where $\alpha_1, \alpha_3$ may be $Y$ or $N$. 

Figure 6.1: Ultimatum Game
We do have other ICE’s, for example, \((99, 1), (Y, N, N)\) and even \((1, 99), (N, N, Y)\), which are also Nash equilibria of this game. However, this game is an extensive game having some information transmission. This suggests an extension of our theory to extensive games or information protocols such as in [9], [10] in order to possibly reduce the set of ICE’s.

In the above ultimatum game, we can allow joint and mutually beneficial deviations. In Fig.6.1, when person \(i\) at role \(a\) deviates, person \(j\) at role \(b\) sees \(i\)’s deviation at role \(a\) before \(j\)’s move. This differs from the strategic game case, where when one person takes an initiative to start a deviation but the other notices it after his move. Thus, in the ultimatum game, person \(i\) could anticipate that \(j\) will respond to \(i\)’s deviation at \(a\) with a mutually beneficial deviation at role \(b\). For example, the deviation from \((99, 1), (Y, N, N)\) to \((50, 50), (Y, Y, Y)\) could be used to eliminate \((99, 1), (Y, N, N)\).

When we allow joint deviations for the ICE, \((50, 50), (\alpha_1, Y, \alpha_3)\) are only the ICE’s.

**Implications of Our Results for Social Morality:** The experimental results often differ from the non-cooperative game-theoretical predictions, but are rather closer to our utilitarian results. Experimental theorists have tried to interpret these 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 specified the basic social context with role-switching, and derived the emergence of cooperation. Our approach may be regarded as providing structural foundations for “fairness”, “altruism”, and “social preferences”.

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. Such behavior might be brought to and observed in experiments.

This gives an experiential grounding for morality, which may be expressed in the form of “utilitarianism” of Theorem 5.3. It has a similarity with Adam Smith’s [19] “moral sentiments” in which a person derives the viewpoint of the (impartial) “spectator” by imagining a social situation. This argument assumes that the spectator has the ability of sympathy and understanding of the target social situation. Our argument explains how the person gets his understanding of the situation and the other’s through role-switching and transpersonal projection.

7. **External and Internal Reciprocities**

Internal reciprocity (2.3), which was used as a key condition in our analysis, represents reciprocity across the roles within the same person \(i\). In this section we show that (2.3) can be motivated and derived from entirely external conditions, by which we mean comparisons of the domains of person \(i\) with those of person \(j\). Also, we give a comment on the relationship between frequency weights and external reciprocities.
Let us start with the accumulated domains $D_1 = (D_{1a}, D_{1b})$ and $D_2 = (D_{2a}, D_{2b})$ for persons 1 and 2 with the regular actions $(s^o_a, s^o_b)$. These domains are externally correlated since the passive experiences of one person are generated by active experiences of the other. Based on this, we could impose the following condition on domains of accumulation: For all $s_r \in S_r$, $r = a, b$ and $i, j = 1, 2$ ($i \neq j$),

$$(s_r; s^{o_r}_{-r}) \in D_{jr(-r)} \text{ implies } (s_r; s^{o_r}_{-r}) \in D_{ir}.$$  \hspace{1cm} (7.1)

That is, if $j$ at role $-r$ keeps a passive experience $(s_r; s^{o_r}_{-r})$, then $i$ keeps the same pair as an active experience. This means that a person is more sensitive to being active than passive. Condition (7.1) has an element of external reciprocity but is a rather weak form since even the non-reciprocal active domains $D^N_1$ and $D^N_2$ given by (2.4) satisfy (7.1).

As time passes, each person may have learned also passive experiences. Eventually, the converse of (7.1) could hold: For all $s_r \in S_r$, $r = a, b$ and $i = 1, 2$,

$$(s_r; s^{o_r}_{-r}) \in D_{j(-r)} \text{ if and only if } (s_r; s^{o_r}_{-r}) \in D_{ir}.$$  \hspace{1cm} (7.2)

The non-reciprocal active domains $D^N_1$ and $D^N_2$ fail to satisfy (7.2), but this does not yet imply the internal reciprocity of (2.3). Condition (7.2) requires the two persons to have the same sensitivities, but allows them to have different trial deviations.

Condition (7.2) describes how one person’s deviation may affect his own and the other’s memories. There is another interpersonal condition on the relationship between both persons’ deviations and memories. It is the coincidence assumption that the two persons have the same trial deviations at each role: for all $s_r \in S_r$, $r = a, b$ and $i = 1, 2$,

$$(s_r; s^{o_r}_{-r}) \in D_{jr} \text{ if and only if } (s_r; s^{o_r}_{-r}) \in D_{ir}.$$  \hspace{1cm} (7.3)

That is, when each person takes the same role, both have the same deviations and memories.

Both (7.2) and (7.3) are external relationships, but they are enough to guarantee the internal reciprocity of (2.3), and also the other external reciprocity.

**Theorem 7.1 (Internal-External Reciprocity).** Conditions (7.2) and (7.3) hold for $D_1 = (D_{1a}, D_{1b})$ and $D_2 = (D_{2a}, D_{2b})$ if and only if (2.3) and

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

**Proof.** When (2.3) and External Reciprocity, the four sets, Proj$(D_{ir})$, $i = 1, 2$ and $r = a, b$ coincide. Hence, the if-part is straightforward. We prove the only-if part. Suppose (7.2) and (7.3) for $D_1$ and $D_2$.

Consider (2.3). Let $(s_a, s_b) \in \text{Proj}(D_{1a})$, i.e., $(s_a, s_b) = (s^o_a, s^o_b)$ or $(s'_a, s'_b)$. First, let $(s_a, s_b) = (s^o_a, s^o_b)$. Then, $(s_a, s^o_b) \in \text{Proj}(D_{2a})$ by (7.3), which is written as $(s'_a; s^o_{-a}) \in \text{Proj}(D_{1a})$.
By (7.2), we have \((s_a; s_{-a}^o) \in \text{Proj}(D_{1(-a)})\), i.e., \((s_a, s_{-a}^o) \in \text{Proj}(D_{1b})\). Next, let \((s_a, s_b) = (s_a^o, s_b)\). Thus, \((s_b; s_{-b}^o) \in \text{Proj}(D_{1(-b)})\). We have \((s_b; s_{-b}^o) \in \text{Proj}(D_{2b})\) by (7.2). Hence, by (7.3), \((s_b; s_{-b}^o) \in \text{Proj}(D_{1b})\). We have shown \(\text{Proj}(D_{1a}) \subseteq \text{Proj}(D_{1b})\). The converse is obtained by a symmetric argument. Thus, we have (2.3).

Consider External Reciprocity. Let \((s_a, s_b) \in \text{Proj}(D_{1a})\), i.e., \((s_a, s_b) = (s_a, s_{-b}^o)\) or \((s_a^o, s_b)\). Let \((s_a, s_b) = (s_a^o, s_b)\). By (7.3), we have \((s_a, s_b^o) \in \text{Proj}(D_{2a})\), i.e., \((s_a; s_{-a}^o) \in \text{Proj}(D_{2a})\). Now, let \((s_a, s_b) = (s_a^o, s_b)\). By (1), \((s_a^o, s_b) \in \text{Proj}(D_{1b})\). This is written as \((s_b; s_{-b}^o) \in \text{Proj}(D_{1b})\). By (7.2), we have \((s_b; s_{-b}^o) \in \text{Proj}(D_{2a})\). We have shown that \(\text{Proj}(D_{1a}) \subseteq \text{Proj}(D_{2a})\). The converse is obtained by a symmetric argument. Thus we have External Reciprocity. ■

We have interpreted frequency weights as a subjective understanding. To study the relationships between these and internal and external reciprocities on \(D_i\)'s, we should refer to objective frequency weights. In the experimental study of Takeuchi et al. [20], the frequency weights are assumed to be externally given as the alternating role-switching as well as no role-switching. Because of the basis of bounded rationality, it would be difficult for a person to evaluate frequency weights accurately. When the objective frequency weights are skewed slightly, a tendency is expected to take them as equally weighted. Nevertheless, when the objective weights are more skewed, the domains \(D_i\)'s could be skewed.

For example, consider the objective frequencies \(\frac{1}{3}\) for role \(a\). Then, person \(i\) experiences role \(b\) twice more than role \(a\). To have the same number of experiences of \(b\), he needs 2 times longer than for \(a\). By Postulate EP1, he may forget previous experiences even if he has had the same number of experiences for role \(a\). Therefore, at a point of time, the domain \(D_{1a}\) may be much smaller than \(D_{1b}\). Thus, the internal reciprocity of (2.3) may not be expected for cases far from objective frequency \(\frac{1}{3}\) for role \(a\).

8. Conclusions

We have introduced the concept of social roles into IGT in order to study an experimental foundation of a person’s and the other’s understanding of their situation. Based on this foundation, we have shown the possibility for the emergence of cooperation, and argued that cooperation is more likely achieved when role-switching is more reciprocal. The foundational study and cooperation result have implications to the three important literatures: (1) Mead’s [15] argument for role-switching and cooperation, (2) cooperative game theory, and (3) noncooperative game theory from the perspective of ex ante decision making. Since our analysis is restricted to the 2-person cases, we first give a comment on this restriction before talking about (1), (2) and (3).

In an extension to situations with more than two persons, we would have a lot of difficulties. We should notice that the number of role assignments is exponentially
increasing with the number of people. Even in the 3-role case, there are many role-assignments. It would be difficult, from the perspective of finite and bounded cognitive abilities, to treat all the role assignments. Role-switching between two persons may be still essential for studying the cases with three or more people.

A key to such an extension is patterned behavior in different but similar situations. An important element of patterned behavior is regularity and uniformity, which could ease difficulties involved in reaching cooperation. This view is related to the very basic presumption of IGT: 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 depicted in Fig. 1.1. This may help us take future steps of extensions of the approach of this paper. Role-switching between two people is a building block for such a situation. In Mead’s [15] baseball example, a pitcher understands a third baseman’s perspective if he has experienced that role a few times, and a catcher understands it also if he plays third, etc. Also, once a pitcher understands the perspective of a third, he may extend his understanding to the other infielders, though some or many details differ from third. Our analysis based on the 2-person cases could be a base for such considerations.

This argument is quite different from the cooperative game theory literature from von Neumann-Morgenstern [21]. The general cooperative game theory starts allowing all possible coalitions to cooperate and giving attainable payoffs by their cooperations. This approach apparently deviates from our basic postulate of people with bounded abilities. However, in the some literature such as that of “an assignment games” initiated by Shapley-Shubik [18], permissible coalitions are restricted to 2-person coalitions. A consideration of a connection to this literature may give a hint to do research in the direction above discussed, though it would be difficult to have a direct connection between this literature and the extension of our approach suggested above.

Finally, we should give comments on (3). Our approach is related to the problem of “common knowledge” or “higher-order beliefs”, though we only informally touch these problems. Often, the common knowledge is regarded as necessary (or sufficient) for the Nash equilibrium concept from the perspective of \textit{ex ante} decision making. Our approach could be regarded as exploring a source for the common knowledge of the game situation, but the reciprocal case, which is central to our approach, allows the cooperation results. That is, in our approach, a kind of “common knowledge” is obtained, and at the same time, cooperation is arising.

This should not be interpreted as meaning that our approach denies the Nash equilibrium from the perspective of \textit{ex ante} decision making based on the common knowledge assumption. The reason is not due to the Nash equilibrium result for the partial use case (Theorem 4.1), but is that we did not take the perspective of \textit{ex ante} decision making for our cooperative result. To have our cooperation result, we needed the whole elements of the dynamic feature of the frequency weights for role-switching and the average payoffs. We have pursued some new possible scenario different from the standard one.
Nonetheless, since our concept of a transpersonal view treats higher-order beliefs, some readers may ask about the relationship of our approach to higher-order beliefs in the game theory literature. Here, we consider two approaches treating higher-order beliefs. One is the universal-type space approach (cf., Mertens-Zamir [16] and Brandenburger-Dekel [2]), and the other is the epistemic logic approach (cf., Fagin, et al. [6], Kaneko [8], and Kaneko-Suzuki [13]). Since the next step of our research is to treat higher-order beliefs more explicitly, it may be helpful to discuss salient differences between our theory and those approaches.

An apparent difference is that our theory asks the source for higher-order beliefs, while the other approaches treat higher-order beliefs as exogenously given. We do not need to discuss this difference furthermore. Rather, it would be helpful to discuss which, the universal-type space approach or the epistemic logic approach, is more natural for an explicit treatment of our approach.

We adopted the representation of “beliefs” in terms of neither types nor subjective probabilities; instead, the beliefs are expressed in terms of classical game theory. The targets of a person’s beliefs are the structures of a game including the regular actions and frequency weights. In the universal-type approach, these are expressed as types; a distinction between two types is basic for the approach, and there is no further structure in a type. We think that the internal structure of an individual view is essential for the present research as well as future developments, since we can talk directly about interpersonal as well as intrapersonal inferences, which are also important aspects of people with bounded abilities.

These structures can be described by a formal language of the epistemic logic approach. This extension has various merits: We can focus on the beliefs about the structure for the persons. This leads us to an explicit treatment of the persons’ logical inferences including inductive and deductive inferences. Also, we may avoid the “common knowledge”; Kaneko-Suzuki [13] already developed an epistemic logic with shallow interpersonal depths. This is also motivated by our basic presumption that people are boundedly rational, as discussed in Section 2.2. Nevertheless, it needs a lot of steps to develop a clear connection between them.

References


