International Game Theory Review

UNDERSTANDING THE OTHER THROUGH SOCIAL ROLES

Mamoru Kaneko

School of Political Science and Economics, Waseda University,
1-6-1, Nishi-Waseda, Shinjuku, Tokyo 169-8050, Japan
mkanekoepi@waseda.jp

J. Jude Kline

School of Economics, University of Queensland,
Brisbane, QLD 4072, Australia
j.kline@uq.edu.au

Received (Day Month Year)
Revised (Day Month Year)

Inductive game theory has been developed to explore the origin of beliefs of a person from his accumulated experiences of a game situation. It 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, especially about the other's payoffs. 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, 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 also consider how a person can use these views for his behavior revision, and we define an equilibrium called an *intrapersonal coordination equilibrium*. Based on this, we show that cooperation will emerge as the degree of reciprocity increases.

Keywords: Inductive Game Theory; Role-Switching; Reciprocity; Strategic Game

Subject Classification: C70, D80.

1. Introduction

The problem of how a person obtains his own beliefs about others' 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 beliefs we refer to are beliefs about the structure of the game including such aspects as 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 others. The present authors, Kaneko and Kline [2008a], Kaneko and Kline [2008b] and Kaneko and Kline [2013], have developed *inductive game theory*

(IGT)[a] in order to explore experiential sources for individual beliefs. For this, the starting assumption of IGT is that a person has little prior information about the game situation in question, which we call the *no prior knowledge assumption*, and we will refer to it from time to time in this paper.

This paper takes one step further from Kaneko and Kline [2008a], Kaneko and Kline [2008b] and Kaneko and Kline [2013] 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. Some behavioral implications follow in addition to the results obtained in Kaneko and Kline [2008a], Kaneko and Kline [2008b] and Kaneko and Kline [2013], which include possible 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 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, they may find a more cooperative approach to family affairs. Although we study a specific 2-person situation with role-switching, it is important to emphasize that each of such situations occurs within a social web like the one described in Fig.1. We use the term "social web" to refer to a variety of social interactions of individuals in society including social gatherings, workplace interactions, and others; not just social interactions conducted on the internet (world wide web).

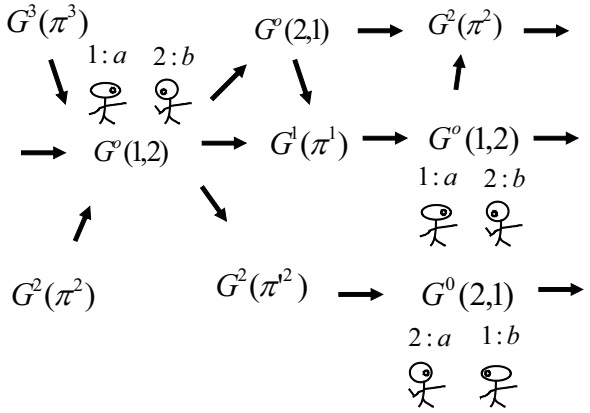

Fig. 1. Social Web

[a]A seminal form of IGT was given in Kaneko and Matsui [1999].

Let us look at this figure in detail in order to elaborate on the development of our theory. First, $G^o(1,2)$ describes an instance of a 2-person game G^o with role a taken by person 1 and role b taken by person 2. The instance $G^o(2,1)$ is based on the same game G^o, except the roles are switched. As mentioned above, we have included different games like $G^2(\pi^2)$ to keep in mind that each player participates in a variety of games, where $G^2(\pi^2)$ is the game G^2 with role assignment π^2.

Our theory does not directly deal with such a social web. However, our various assumptions based on "bounded rationality" are better understood by looking at the social web rather than a recurrence of a particular game. Since the entire social web exceeds the reasonable capacity of conscious considerations of each instance and occurrence of a game situation, people are rule-governed with their behavior, collections of experiences, understanding, and behavioral revisions. This presumption of a social web and bounded rationality will be used to motivate various developments of our concepts.

We distinguish between a role and a person to describe the role-switching. A role is a position in the game, and a person, on the other hand, is an individual who participates in a social web. The concept of a player in the standard sense consists of a role and a person in this paper. 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^o with roles a and b. We presume that each person $i = 1, 2$ can separate this G^o from the other games, and he keeps memories from playing G^o, in the form of a *memory kit*[b] to be described in Section 2. The accumulated memories in the memory kit substitute for the no prior knowledge assumption. He uses his memories over the repeated plays of G^o to both construct his view of G^o and to adjust his behavior in G^o. We are interested in how role-switching affects his view and behavior.

Our development and findings are influenced by the works of Mead [1934] (cf., Collins [1988], Chap.7) on the importance of role-switching for obtaining a social view, and of Lewis [1969] on common knowledge. When two people switch social roles reciprocally, each person has seen the other 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 give each person "reason to believe" that the other has had similar experiences. This is similar to Lewis's requirement of "reason to believe" in his definition of common knowledge[c], though we treat "shallow interpersonal beliefs" in a non-formalized manner in this paper.

Mead [1934] 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

[b]The cognitive limititions on a person are implicit in our formulation of a memory kit in that it ignores the precise sequence of past plays of G^o. This formulation is justified by the epistemic postulates formulated in Section 2.2 which embody bounded rationality aspects of people.

[c]In the more recent literature of epistemic logic, the fixed-point characterization of "common knowledge" has an aspect similar to "reason to believe" (cf., Fagin *et al.* [1995] and Kaneko [2002]).

was emphasized by Mead [1934] and his predecessor, Cooley [1902], to argue the pervasiveness of cooperation in human society. This was criticized as too naive by later sociologists (see Collins [1988], Chap.7). In our theory, cooperation is one possibility obtained by role-switching, but not necessarily guaranteed. On a purely motivational level, our work is also related to the work of Smith [1759] on mutual sympathy, and to the team reasoning approach of Bacharach [1999], which will be discussed in Sections 6 and 8.

Here, we give a summary of the new concepts to be used in this paper, while emphasizing some results. Since our approach is based on IGT developed in Kaneko and Kline [2008a], the concepts of a memory kit and inductively derived views (i.d.views) defined there will be taken and adjusted to fit the current context of role-switching.

The new concepts introduced in this paper are:

(A): A 2-role Game in a Recurrent Situation with Role-switching: A particular situation G^o in Fig.1 is given as a 2-role strategic game G^o. Each person takes one role in each occurrence of G^o, and the persons switch the roles from time to time. Frequency of each role is externally given, but each person has some subjective perception of this frequency. The key departure from the previous work in IGT is the introduction of role-switching.

In addition to role-switching, each person makes trials/errors, which take the form of temporary deviations from the regular actions. These give new experiences other than what each is regularly experiencing. If a person has such experiences with some frequency, he may keep and accumulate them. Thus, each accumulates his experiences up to his cognitive abilities, which are summarized as the concept called a *memory kit* κ_i. The step of trials/errors is informal in this paper, and will

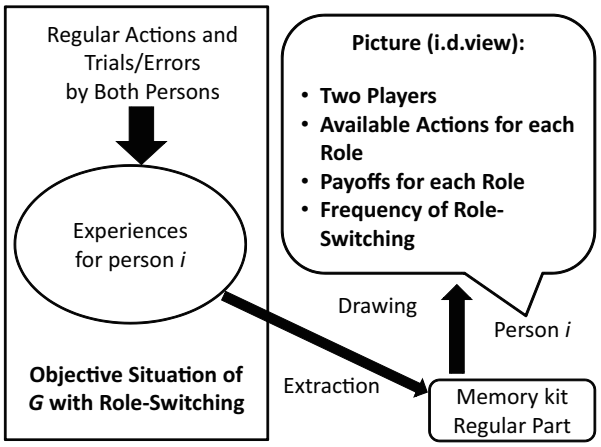

Fig. 2. Extraction and Drawing

Direct Understanding				Transpersonal Understanding
g^{ii}	$\longleftarrow$ static	*memory kit* κ_i	static $\longrightarrow$	g^{ij}
$\downarrow$ Γ^{ii}		adding temporal elements		$\downarrow$ Γ^{ij}

Fig. 3.

Table 1. PD

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

be discussed in Section 2.2.

In Fig.2, person i obtains a memory kit by extracting regularities hidden in experiences, and then draws a picture of the situation based on the memory kit. This picture may give new behavioral implications. Those will be discussed in (B) and (C).

(B): Direct and Transpersonal Understandings: Fig.3 shows a map on connections among these components and the memory kit. The direct understanding g^{ii} of G^o describes person i's own understanding of the game G^o based on his memory kit purely from person i's perspective. Conceptually, this is the i.d.view in the sense of Kaneko and Kline [2008a]. However, experiences with each role enables person i to infer the payoffs for both roles under certain postulates.

The transpersonal understanding g^{ij} is person i's thought of j's understanding of G^o. We require this g^{ij} to be derived from person i's memory kit. For this, person i projects his direct understanding g^{ii} to person j, but this projection is constrained by "reason to believe" mentioned above. The term "transpersonal" reflects the postulates of projection and "reason to believe".

Both understandings g^{ii} and g^{ij} are still static descriptions of G^o, though the situation itself is recurrent described in Fig.1. We add, to g^{ii} and g^{ij}, his subjective perception of frequency of role-switching as an temporal element, and the resulting understandings are represented by Γ^{ii} and Γ^{ij} for his own and the other's understandings, respectively. These form an inductively derived view $(\Gamma^{ii}, \Gamma^{ij})$, which is an extension of an i.d.view given in Kaneko and Kline [2008a] in the present context.

(C): Intrapersonal Coordination Equilibrium (ICE): The equilibrium we call an ICE is based on Γ^{ii} and Γ^{ij} together with some specific conjectural postulate. This is called the *full use* and will be discussed in Section 5.1. The term "intrapersonal" refers to the fact that this occurs in the mind of person i. When role-switching is sufficiently reciprocal, the detailed difference in the subjective frequency weights of

Table 2. SH1

	$\mathbf{s}_{b1}$	$\mathbf{s}_{b2}$
$\mathbf{s}_{a1}$	$(7,7)^{NE,ICE}$	$(1,4)$
$\mathbf{s}_{a2}$	$(4,1)$	$(2,2)^{NE}$

Table 3. SH2

	$\mathbf{s}_{b1}$	$\mathbf{s}_{b2}$
$\mathbf{s}_{a1}$	$(7,7)^{NE,ICE}$	$(1,4)$
$\mathbf{s}_{a2}$	$(4,1)$	$(3,3)^{NE,ICE}$

role-switching effectively disappears, which will be stated in the *utilitarian theorem* (Theorem 3): ICE is determined by the unweighted joint payoff-sum maximization on the domain over unilateral deviations. We also consider the *partial use* case, in which only the direct understanding H^{ii} together with a certain conjectural postulate will be used, and the resulting outcome is a Nash equilibrium.

Here, we illustrate the utilitarian theorem in three examples. In Table 1 (Prisoner's Dilemma), the unique ICE with sufficient role-switching is given as $(\mathbf{s}_{a1}, \mathbf{s}_{b1})$, which is not an NE. Table 2 (Stag Hunt 1) also has $(\mathbf{s}_{a1}, \mathbf{s}_{b1})$ as the unique ICE with sufficient role switching, which is now a Nash equilibrium, but not the only one. In Table 3 (Stag Hunt 2), the other Nash equilibrium $(\mathbf{s}_{a2}, \mathbf{s}_{b2})$ becomes a new ICE.

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 predict the utilitarian result of joint payoff-sum maximization. Thus, we predict the emergence of cooperation from sufficient role-switching. This has some implications for the theory of social morality and also for the literature of cooperative game theory, which will be discussed in Sections 6 and 8.

Here, we should give a comment on the *repeated game approach* (cf., Osborne-Rubinstein Osborne and Rubinstein [1994]), since it appears to be related in that both treat recurrent situations. Nevertheless, these approaches have radical differences in their aims and basic assumptions. As emphasized above, IGT primarily aims to study the origin and emergence of the personal understandings of the game structure, and behavioral consequences are somewhat secondary. It is possible to introduce "social roles" to the repeated game approach, but our primary question about the other person's understanding of the game is simply irrelevant, since it assumes that the entire situation is known to the players as a huge one-shot game. Still, it may be useful to state what would be different in the resulting outcomes with role-switching between our theory and the repeated game approach. In the latter, ICE could be sustained by an equilibrium, but many more outcomes are expected as suggested by the Folk Theorem. One point of our theory is to give more precise predictions like the utilitarian result.

The paper is written as follows. Section 2 gives the basic definitions of a 2-role game and domains of experiences. Section 3 defines a 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. Section 5 formulates an intrapersonal coordination equilibrium. 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 and some 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 a 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.

2.1. *2-Role Strategic Games, 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*, $S_r = \{\mathbf{s}_{r1}, ..., \mathbf{s}_{r\ell_r}\}$ is a finite set of *actions,* and $h_r : S_a \times S_b \to \mathbf{R}$ is a *payoff function* for role $r = a, b$. Each role is taken by a *person* $i = 1, 2$. A *role assignment* π is a one-one mapping from $\{a, b\}$ to $\{1, 2\}$, which describes the role taken by each person in a particular occurrence 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* $G(\pi)$ is given by adding a role assignment $\pi = (i_a, i_b)$ to a 2-role strategic game G.

When a game G occurs assigning person i to a role r, i.e., $\pi(r) = i$, he is told that the available actions are given by S_r. Also, he has now the observations, which are explicitly stated by the following.

Assumption Ob: After each play of G, his observations are summarized by the current role r, the action pair (s_a, s_b) played and his own payoff value $h_r(s_a, s_b)$ in $G(\pi)$.

Since payoffs are subjective elements, each person is assumed to observe only his own payoff value in each play. Here, 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.

Each person has made trials/errors to generate experiences and accumulate them through his limited cognitive abilities, the process of which will be informally discussed in Section 2.2. Here, we start with a summary of such accumulated experiences, which is given as a *memory kit* $\kappa_i = \langle (s_a^o, s_b^o), (D_{ia}, D_{ib}), (h_{ia}, h_{ib}); (\rho_{ia}, \rho_{ib}) \rangle$ consisting of:

κ1: the pair $(s_a^o, s_b^o) \in S_a \times S_b$ of *regular actions*;

κ2: the *accumulated domain of experiences* $D_i = (D_{ia}, D_{ib})$ from taking role $r = a, b$, where $(s_a^o, s_b^o) \in D_{ia} \cup D_{ib} \subseteq S_a \times S_b$;

$\kappa 3$: the *observed payoff functions* (h_{ia}, h_{ib}) over D_i, where $h_{ir} : D_{ir} \to \mathbf{R}$ and $h_{ir}(s_a, s_b) = h_r(s_a, s_b)$ for all $(s_a, s_b) \in D_{ir}$ and $r = a, b$;

$\kappa 4$: the (perceived) *frequency weights* (ρ_{ia}, ρ_{ib}) for roles a and b, where $\rho_{ia}, \rho_{ib} \geq 0$ and $\rho_{ia} + \rho_{ib} = 1$.

The regular actions s_a^o and s_b^o defined in condition $\kappa 1$ are part of the memory kit. The players are presumed to know that these actions are regularly played by the two people. These are discussed in detail in Postulate BH1 of Section 2.2. Condition $\kappa 2$ 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 $(s_a^o, s_b^o) \in D_{ia} \cup D_{ib}$. 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 (ρ_{ia}, ρ_{ib}) expresses person i's perceived frequencies of roles a and b.

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.

We assume the following on a memory kit: for all $r = a, b$,

$$\text{if } (s_a, s_b) \in D_{ir}, \text{then } (s_a, s_b^o) \in D_{ir} \text{ and } (s_a^o, s_b) \in D_{ir}; \tag{1}$$

$$\rho_{ir} = 0 \text{ if and only if } D_{ir} = \emptyset. \tag{2}$$

Condition (1) states that if (s_a, s_b) is accumulated in D_{ir}, then (s_a, s_b^o) and (s_a^o, s_b) coming from the unilateral trials of s_a and s_b from (s_a^o, s_b^o) are also accumulated in $D_{ia} \cup D_{ib}$. Condition (2) states that if $\rho_{ir} = 0$, person i has no recollection of being in role r, and *vice versa.* Using (1), we have the following lemma.

Lemma 1. *$D_{ir} \neq \emptyset$ if and only if $(s_a^o, s_b^o) \in D_{ir}$; equivalently, $D_{ir} = \emptyset$ if and only if $(s_a^o, s_b^o) \notin D_{ir}$.*

When $(s_r; s_{-r}^o) \in D_{ir}$, it is called an *active experience* for person i at role r, since person i's own deviation causes this experience. When $(s_r; s_{-r}^o) \in D_{i(-r)}$, it is a *passive experience* for person i at role $-r$, since this experience for i comes from j's deviation.

Reciprocity is important in this paper, but we have various degrees of it. First, we define the set $\text{Proj}(T) := \{(s_a, s_b) \in T : s_a = s_a^o \text{ or } s_b = s_b^o\}$, where $T \subseteq S_a \times S_b$. Then, we say that D_{ia} and D_{ib} are *internally reciprocal* iff

$$\text{Proj}(D_{ia}) = \text{Proj}(D_{ib}). \tag{3}$$

This requires the equivalence of D_{ia} and D_{ib} up to unilateral changes from the regular actions (s_a^o, s_b^o) : Indeed, (3) implies $D_{ia} \neq \emptyset$ and $D_{ib} \neq \emptyset$, and we have $(s_a^o, s_b^o) \in D_{ia} \cap D_{ib}$ by Lemma 1; (3) is the equivalence up to unilateral changes from (s_a^o, s_b^o). A stronger reciprocity is $D_{ia} = D_{ib}$, but (3) is more relevant in this paper. The term "internally" stresses that it is about reciprocity across domains

of a single person i. Since, however, person i interacts with person j, this internal reciprocity may be externally motivated. In Section 7, we show how (3) may be derived from external conditions on reciprocity.

By (2), (3) is impossible when $\rho_r = 0$. It would be natural to introduce lower and upper bounds on ρ_r for internal reciprocity to hold. In Section 5.2, we will give bounds when we talk about the utilitarian result (Theorem 3).

Here we give two examples of domains to discuss internal reciprocity.

(1)(Non-reciprocal Active Domain): Let $D_1^N = (D_{1a}^N, D_{1b}^N)$ be given as follows:

$$D_{1a}^N = \{(s_a, s_b^o) : s_a \in S_a\} \text{ and } D_{1b}^N = \emptyset. \tag{4}$$

Let $D_2^N = (D_{2a}^N, D_{2b}^N)$ be defined in the symmetric manner. By $\kappa 4$ and (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. Here, internal reciprocity (3) does not hold.

There are other non-reciprocal domains, even ones where each person is sensitive to active and passive deviations. We have also varieties of reciprocal domains. We focus on one:

(2)(Active-Passive Domain): $D_1^{AP} = (D_{1a}^{AP}, D_{1b}^{AP})$ for person 1 is given as:

$$D_{1a}^{AP} = D_{1b}^{AP} = \{(s_a, s_b^o) : s_a \in S_a\} \cup \{(s_a^o, s_b) : s_b \in S_b\}. \tag{5}$$

This domain for person 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 to joint-trials.

2.2. *Informal Postulates for Behavior and Accumulation of Memories*

Our mathematical theory starts with a memory kit $\kappa_i = \langle (s_a^o, s_b^o), (D_{ia}, D_{ib}), (h_{ia}, h_{ib}); (\rho_{ia}, \rho_{ib}) \rangle$. A memory kit κ_i is extracted from accumulated memories from experiences. We explain the basic postulates given in Kaneko and Kline [2008a] and some additional postulate in the present context[d].

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

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.

[d]We give various postulates, under which our theory will be built. Although they look natural, we do not claim that they are the absolute starting assumptions. Rather, we start the present theory with those "postulates", but they are open to further theoretical and/or experimental tests.

The role switching postulate BH0 is newly introduced in this paper. As mentioned in Section 1, it gives a source for a person's view of the other. Postulates BH1 and BH2 are about the rule-governed behavior and trial deviations.

Postulate BH1 states that the players follow some regular behavior in the recurrent situation. In the beginning, each person may have 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; perhaps he found it worked well in the past or he was taught to follow it. After a person accumulates enough memories and forms a view on the social situation, the regular behavior may represent a conscious decision.

A general form of BH1 is that each person's behavior follows a possibly more complex pattern including dependence upon a role assignment. This regularity is emphasized in the definition of "convention" due to Lewis [1969]. While IGT shares bases with his theory, it also targets the experiential sources for persons' beliefs from regularity in behavior and also trials/errors stated in BH2. Here, we adopt the above form of BH1 to keep our developments simple, which postulates that it depends upon some specified regular action s_r^o.

Postulate BH2 describes how a person makes trials and errors. 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.

When trials/errors in BH2 are added, the regular behavior postulated in BH1 manifests its main function in this paper: Learning is possible only when there is some regularity identifiable to the learner. In contrast, if the people behave in arbitrary ways (randomly, or following complicated patterns), a person who has made a trial deviation cannot find causality up to his cognitive ability. Thus, simpler regularity is more important, and BH1 assumes the simplest form.

Person i's local (short-term) memory is what he receives in an instant, which takes the form of $\langle r, (s_a, s_b), h_r(s_a, s_b)\rangle$ by assumption Ob. On the other hand, the accumulated, i.e., *long-term*, memories are described by D_{ir}. Once the triple $\langle r, (s_a, s_b), h_r(s_a, s_b)\rangle$ is newly transformed to a long-term memory, his domain D_{ir} is extended to $D_{ir} \cup \{(s_a, s_b)\}$, and "$h_r(s_a, s_b)$" is recorded in the memory kit κ_i. For the transition from local to long-term memories, we have various scenarios. Here we list postulates based on bounded memory abilities which restrict those scenarios.

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, which is a justification for $\kappa 2$. A pair obtained by only one person's deviation from the regular behavior is more likely to remain in his memory than pairs obtained by joint deviations, which supports condition (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 involves a strong memory ability, which is incompatible with EP1 and EP2. We assume that the long-term memories accumulated from his experiences take the form of a memory kit as described by $\kappa 1 - \kappa 4$. An explicit process of transformation from experiences (short-term memories) to long-term memories is given in Akiyama *et al.* [2013]. Also, those postulates can be tested in experiments; Takeuchi *et al.* [2013] 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

Person i has been playing the game situation G with role-switching and has accumulated experiences, which are summarized as a memory kit $\kappa_i = \langle (s_a^o, s_b^o), (D_{ia}, D_{ib}), (h_{ia}, h_{ib}); (\rho_{ia}, \rho_{ib}) \rangle$. Now, he may question what situation he has been in, and what the other thinks. Here, we let i consider the constant part, i.e., G, of the situation from his memory kit κ_i, and postpone his consideration of temporal elements to Section 4. In terms of Fig.2, he now constructs his direct understanding g^{ii} and transpersonal understanding g^{ij} in a static manner.

Now, person i constructs, based on his memory kit κ_i:

(DU) his own (direct) understanding about the game structure G;

(TP) his (transpersonal) thought of the other's understanding about G.

The former is simply obtained by combining his experiences, while the latter needs some additional interpersonal thinking. In this section, we describe how the person deals with these problems.

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

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

This will be presently formulated as his own understanding g^{ii} of game G. Let us turn our attention to his thought of the other's understanding g^{ij} of G. Under assumption Ob, it suffices to consider the subjective elements, i.e., payoffs, for the other person. 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.

Assumption Ob states that he observes only his own payoffs. To think about the other's, he needs to use his own 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 need to have some experiential evidence to believes that j has the same experienced payoffs as what i has.

A simple metaphor may help the reader understand those postulates[e]: A boy notices that a girl appears to be suffering the agony of a broken heart. In addition, it is assumed that he has experienced a broken heart as well as has some experiential source indicating her broken heart. By these, he is able to project his former feelings onto her. The ability of projecting his former experiences is stated by TP1, and the reason to believe her heart is broken is required by TP2. In the case where he has no source for her broken heart, he may doubt her behavior; he may think that she is simply pretending.

Let us return to our mathematical world. Now, person i constructs, based on his memory kit $\kappa_i = \langle (s_a^o, s_b^o), (D_{ia}, D_{ib}), (h_{ia}, h_{ib}); (\rho_{ia}, \rho_{ib}) \rangle$, his *direct understanding* and also infers/guesses his *transpersonal understanding* of the other's understanding, which are the left-hand and right-hand columns of Fig.3. We prepare a unexperiential payoff value $\theta_r^i(s_a, s_b)$ attached to $(s_a, s_b) \in S_a^i \times S_b^i$. In (1) of Definition 1, $\theta_r^i(s_a, s_b)$ is used only for the part $S_a^i \times S_b^i - D_{ir}$, and in (2), it is used for $S_a^i \times S_b^i - D_{ia} \cap D_{ib}$.

They are formulated in the following definition.

Definition 1 (Direct and Transpersonal Understandings).

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

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

ID2^{ii}*: for* $r = a, b$, $h_r^{ii} : S_a^i \times S_b^i \to \mathbf{R}$ *is defined as follows:*

$$h_r^{ii}(s_a, s_b) = \begin{cases} h_r(s_a, s_b) \text{ if } (s_a, s_b) \in D_{ir} \\ \\ \theta_r^i(s_a, s_b) \quad \text{otherwise.} \end{cases} \tag{6}$$

(2) The transpersonal understanding *(*tp-*understanding)* from κ_i *by person* i *for person* j *is given as* $g^{ij} = (a, b, S_a^i, S_b^i, h_a^{ij}, h_b^{ij})$, *where* h_a^{ij} *and* h_b^{ij} *are new and given as:*

[e]In the example of a base ball team, Mead [1934] argued that switching positions often help the players to improve their performance.

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

$$h^{ij}_r(s_a, s_b) = \begin{cases} h_r(s_a, s_b) \text{ if } (s_a, s_b) \in D_{ir} \cap D_{i(-r)} \\ \\ \theta^i_r(s_a, s_b) \quad \text{otherwise.} \end{cases} \tag{7}$$

The components of g^{ii} and g^{ij}, except θ^i_r for the unexperienced part of $S^i_a \times S^i_b$, are determined by memory kit κ_i. 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 tp-understanding g^{ij} differs from g^{ii} only in its definition (7) of the payoff function. In (6), the if-clause is "$(s_a, s_b) \in D_{ir}$" , while it is "$(s_a, s_b) \in D_{ir} \cap D_{i(-r)}$" in (7). That is, for person i's own understanding to have the payoff $h_r(s_a, s_b)$, he is simply required to have (s_a, s_b) in his accumulated domain D_{ir}. On the other hand, the latter requires him to have (s_a, s_b) in both D_{ir} and $D_{i(-r)}$, i.e., to have experienced it at both roles. In this case, person i has both experienced the payoff $h_r(s_a, s_b)$ at role r, and has reason to believe person j experienced it when j was at role r. Hence, person i projects this payoff onto person j.

The value $\theta^i_r(s_a, s_b)$ expresses an unknown (unexperienced) payoff, which is used for his own and the other's understanding. So far, we have not put any restriction on these values. Now, we give some restrictions on the values. We give two conditions.

The first applies to an experienced role r.

(**ER**): If $\rho_{ir} \neq 0$, i.e., $(s^o_a, s^o_b) \in D_{ir}$, then:

$$\theta^i_r(s_a, s_b) < h_r(s^o_a, s^o_b) \text{ for all } (s_a, s_b) \in \text{Proj}(S^i_a \times S^i_b). \tag{8}$$

Here, the experienced payoff $h_r(s^o_a, s^o_b)$ is the reference point, and (8) requires this regular payoff to be larger than any unexperienced imagined value assigned to a unilateral deviation from the regular behavior. The condition will be used in Theorem 2.(1).

The second condition applies to an unexperienced role r. In this case, even the regular pair is unexperienced by i at role r.

(**UR**): If $\rho_{ir} = 0$, i.e., $(s^o_a, s^o_b) \notin D_{ir}$, then:

$$\theta^i_r(s_a, s_b) = \theta^i_r(s^o_a, s^o_b) \text{ for all } (s_a, s_b) \in \text{Proj}(S^i_a \times S^i_b). \tag{9}$$

That is, the payoffs for the unexperienced role are assumed to be constant, though they are used in two different ways: Since $(s_r; s^o_{-r}) \notin D_{ir}$, $\theta^i_r(s_r; s^o_{-r})$ is used for the imagined $h^{ii}_r(s_r; s^o_{-r})$ and for $h^{ij}_r(s_r; s^o_{-r})$. Since $(s^o_r; s_{-r}) \in D_{i(-r)}$, we have $h^{ii}_{-r}(s^o_r; s_{-r}) = h_{-r}(s^o_r; s_{-r})$ but $h^{ij}_r(s^o_r; s_{-r}) = \theta^i_r(s^o_r; s_{-r})$; that is, $\theta^i_r(s^o_r; s_{-r})$ is used for person i's thinking of j's payoffs. This condition will be used in Theorems 1 and 2.(2).

Let us exemplify the above definitions with the PD game of Table 1 assuming the regular actions $(s^o_a, s^o_b) = (\mathbf{s}_{a1}, \mathbf{s}_{b1})$: In those examples, for simplicity, $\theta^i_r(s_a, s_b)$ is assumed to depend upon only a role, so we write simply θ_a and θ_b for unexperienced payoffs.

Table 4. g^{11}

	$\mathbf{s}_{b1}$
$\mathbf{s}_{a1}$	$(5, \theta_b)$
$\mathbf{s}_{a2}$	$(6, \theta_b)$

Table 5. g^{12}

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

Table 6. g^{11} and g^{12}

$a\backslash b$	$\mathbf{s}_{b1}$	$\mathbf{s}_{b2}$
$\mathbf{s}_{a1}$	$(5, 5)$	$(1, 6)$
$\mathbf{s}_{a2}$	$(6, 1)$	(θ_a, θ_b)

(1)(Non-reciprocal Active Domain): Let (D^N_{1a}, D^N_{1b}) be given as the non-reciprocal domain of (4); the situation consists of repetitions of $G(1, 2)$. Person 1's d-understanding $g^{11} = (a, b, S^1_a, S^1_b, h^{11}_a, h^{11}_b)$ is given as: $S^1_a = \{\mathbf{s}_{a1}, \mathbf{s}_{a2}\}$ and $S^1_b = \{\mathbf{s}_{b1}\}$. Since 1 has experiences only at role a, the payoffs $(h^{11}_a(s_a, s_b), h^{11}_b(s_a, s_b))$ are given in Table 4.

His tp-understanding g^{12} differs from g^{11} only in the payoff functions h^{12}_a, h^{12}_b, which are described as Table 5. Since 1 has experienced only role a, he cannot infer the payoff value for role b. Although he is sure that 2 is enjoying only role b, 1 has no experiences for $h_b(s_a, s_b)$ and puts $h^{12}_b(s_a, s_b) = \theta_b$. Then, 1 is sure that 2 has no experiences for $h_a(s_a, s_b)$, which implies $h^{12}_a(s_a, s_b) = \theta_a$. Thus, g^{12} is summarized as Table 5.

The above observations hold more generally. Let $g^{ii} = (a, b, S^i_a, S^i_b, h^{ii}_a, h^{ii}_b)$ and $g^{ij} = (a, b, S^i_a, S^i_b, h^{ij}_a, h^{ij}_b)$ be the d- and tp-understandings. Here, neither (8) nor (9) are used.

Lemma 2. *Let $\rho_{ir} = 1$. Then, $h^{ii}_{-r}(s_a, s_b) = \theta^i_{-r}(s_a, s_b), h^{ij}_r(s_a, s_b) = \theta^i_r(s_a, s_b)$ and $h^{ij}_{-r}(s_a, s_b) = \theta^i_{-r}(s_a, s_b)$ 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). By (6) and (7), we have the stated equations. □

(2):(Reciprocal Active-Passive Domain): Let $D^{AP}_1 = (D^{AP}_{1a}, D^{AP}_{1b})$ be the domains given by (5). Then, $S^1_a = \{\mathbf{s}_{a1}, \mathbf{s}_{a2}\}$ and $S^1_b = \{\mathbf{s}_{b1}, \mathbf{s}_{b2}\}$. Both g^{11} and g^{12} are given by Table 6. 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.

Internal reciprocity will be important in our later analysis. We note the following lemma.

Lemma 3. *Suppose internal reciprocity (3). Then, g^{ii} coincides with g^{ij} up to the active/passive experiences, i.e., for all $(s_a, s_b) \in \mathrm{Proj}(S^i_a \times S^i_b)$ and $r = a, b$,*

$$h^{ii}_r(s_a, s_b) = h^{ij}_r(s_a, s_b) = h_r(s_a, s_b). \tag{10}$$

Proof. Suppose (3), i.e., $\mathrm{Proj}(D_{ia}) = \mathrm{Proj}(D_{ib})$. Then, we show $\mathrm{Proj}(S^i_a \times S^i_b) = \mathrm{Proj}(D_{ia}) = \mathrm{Proj}(D_{ib})$. Let $(s_a, s_b) \in \mathrm{Proj}(S^i_a \times S^i_b)$. 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 (1), $(s_a, s^o_b) \in D_{ia}$ or $(s_a, s^o_b) \in D_{ib}$. Since $\mathrm{Proj}(D_{ia}) = \mathrm{Proj}(D_{ib})$, we have $(s_a, s^o_b) \in \mathrm{Proj}(D_{ia}) = \mathrm{Proj}(D_{ib})$. In the second case, we have also $(s^o_a, s_b) \in \mathrm{Proj}(D_{ia}) = \mathrm{Proj}(D_{ib})$. Conversely, let $(s_a, s_b) \in \mathrm{Proj}(D_{ia}) = \mathrm{Proj}(D_{ib})$. Then, $(s_a, s_b) \in D_{ia}$, which implies $s_a \in S^i_a$. Similarly, $s_b \in S^i_b$. Thus, $(s_a, s_b) \in S^i_a \times S^i_b$. Since $s_a = s^o_a$ or $s_b = s^o_b$, we have $(s_a, s_b) \in \mathrm{Proj}(S^i_a \times S^i_b)$. □

4. Inductively Derived View $\Gamma^i = (\Gamma^{ii}, \Gamma^{ij})$ for Role-Switching

We extracted the regular part from the memory kit κ_i for the understandings g^{ii} and g^{ij}. The situation also includes the temporal aspects of the frequency weights (ρ_{ia}, ρ_{ib}) and of trials/errors from the regular actions. We regard trials/errors as temporary flux. By taking only the frequency weights, we extend g^{ii} and g^{ij} into Γ^{ii} and Γ^{ij}, as illustrated in Fig.3. Then we consider the stability of the regular actions (s^o_a, s^o_b) against behavioral revision based on $\Gamma^i = (\Gamma^{ii}, \Gamma^{ij})$. In this section, we consider the partial use of Γ^i, i.e., only Γ^{ii}, for stability. The full use of Γ^i will be discussed in Section 5.

4.1. *Inductively Derived Views*

The *d*- and *tp*-understandings g^{ii} and g^{ij} include separate payoff functions h^{ii}_a, h^{ii}_b and h^{ij}_a, h^{ij}_b for roles a and b. When person i takes the frequency weights (ρ_{ia}, ρ_{ib}) into account, he combines those payoff functions h^{ii}_a, h^{ii}_b and h^{ij}_a, h^{ij}_b to the temporal payoff functions H^{ii} and H^{ij}. These are the main parts of the temporal *d*- and *tp*-understandings Γ^{ii} and Γ^{ij}.

To describe these H^{ii} and H^{ij} formally, we introduce the expression $(s_a, s_b)_r$ meaning that person i at role r plays s_r while j at role $-r$ plays s_{-r}. Here, to avoid a confusion, we use different strategy pairs (s_a, s_b) and (t_a, t_b), which may be identical. We consider the recurrent situation where $(s_a, s_b)_a$ is played with frequency ρ_{ia}, and $(t_a, t_b)_b$ is played with frequency $\rho_{ib} = 1 - \rho_{ia}$.

The weighted payoff functions H^{ii} and H^{ij} are now defined as follows:

$$H^{ii}((s_a, s_b)_a, (t_a, t_b)_b) = \rho_{ia} h^{ii}_a(s_a, s_b) + \rho_{ib} h^{ii}_b(t_a, t_b); \tag{11}$$

$$H^{ij}((s_a, s_b)_a, (t_a, t_b)_b) = \rho_{ia} h^{ij}_b(s_a, s_b) + \rho_{ib} h^{ij}_a(t_a, t_b). \tag{12}$$

In the right-hand side of (12), weights ρ_{ia} and ρ_{ib} are associated with $h^{ij}_b(s_a, s_b)$ and $h^{ij}_a(t_a, t_b)$, respectively, since H^{ij} is the payoff function considered for person

j in the mind of person i[f]. We take the domain $(S_a^i \times S_b^i)^2$ for H^{ii} and H^{ij} for simplicity.

The above definitions may look very restrictive relative to the standard definition of the evaluation of an infinite stream of outcomes in the repeated game approach (cf., Osborne and Rubinstein [1994]). Nevertheless, their restrictiveness is rather the point and is 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 2. *The* temporal d-understanding Γ^{ii} *and* temporal tp-understanding Γ^{ij} *from the memory kit* κ_i *are given as*

$$\Gamma^{ii} = \langle (s_a^o, s_b^o), (S_a^i, S_b^i), (\rho_{ia}, \rho_{ib}), H^{ii} \rangle; \text{ and } \Gamma^{ij} = \langle (s_a^o, s_b^o), (S_a^i, S_b^i), (\rho_{ia}, \rho_{ib}), H^{ij} \rangle.$$

The pair $\Gamma^i = (\Gamma^{ii}, \Gamma^{ij})$ *is the* inductively derived view *(*i.d.view*) for person* i *derived from the memory kit* κ_i*. We abbreviate "temporal" for* Γ^{ii} *and* Γ^{ij} *when no confusions with* g^{ii} *and* g^{ij} *are caused.*

This definition has various differences from those given in Kaneko and Kline [2008a], Kaneko and Kline [2008b] and Kaneko and Kline [2013]. One apparent difference is that it is given to a strategic game but not an extensive game (or an information protocol). More importantly, the inclusion of frequency weights is crucial and new. Perhaps, we should remark that the determination of an i.d.view $\Gamma^i = (\Gamma^{ii}, \Gamma^{ij})$ by a memory kit κ_i is unique, except for that the values (θ_a^i, θ_b^i) are not determined by experiences:

$$\kappa_i \mapsto \Gamma^i = (\Gamma^{ii}, \Gamma^{ij}) \tag{13}$$

The unique determination of an i.d.view holds also for the definitions in Kaneko and Kline [2008a] and Kaneko and Kline [2008b], but not in Kaneko and Kline [2013]. In Kaneko and Kline [2013], each memory itself may contain some partiality, which allows for a multiplicity of i.d.views and affects even i's decision making.

The *d*-understanding Γ^{ii} may differ from the *tp*-understanding Γ^{ij} only in their payoffs; still, we include the other components to make the structures of Γ^{ii} and Γ^{ij} explicit. The *tp*-understanding Γ^{ij} can be used for person i's prediction about j's decision making. However, it remains possible for person i to ignore it by using some trivial prediction. We divide our analysis into two cases:

C0(Partial Use): Person i uses only the *d*-understanding Γ^{ii}.

C1(Full Use): Person i uses not only Γ^{ii} but also the *tp*-understanding Γ^{ij}, in order to predict how person j will react.

[f]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 and Morgenstern [1944]. See Hu [2013] for a direct approach to expected utility theory from the frequentist perspective.

Table 7. C0; Partial Use

Γ^{ii} +		
conjectural postulate $(*)$	$\Rightarrow$	Partial-Use Equilibrium

Table 8. C1; Full Use

$(\Gamma^{ii}, \Gamma^{ij})$ +		
conjectural postulate $(**)$	$\Rightarrow$	Utilitarian Theorem

We handle C0 in Section 4.2, where it is shown that C0 together with the conjectural postulate $(*)$ leads to Nash equilibrium type behavior. The case C1 will be 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, where we will add the conjectural postulate $(**)$, different from $(*)$.

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

In the case C0, we adopt the following conjectural postulate by person i when he takes an intensional deviation:

$(*)$: person j sticks to the regular actions (s_a^o, s_b^o).

Then, person i may choose a maximum point in S_r^i against the regular action s_{-r}^o. We require that the regular actions (s_a^o, s_b^o) be free from such behavioral revisions. Thus, we have the following definition: (s_a^o, s_b^o) is a *partial-use equilibrium (PUE)* in Γ^{ii} iff for all $s_a \in S_a^i$ and $s_b \in S_b^i$,

$$H^{ii}((s_a^o, s_b^o)_a, (s_a^o, s_b^o)_b) \geq H^{ii}((s_a, s_b^o)_a, (s_a^o, s_b)_b). \tag{14}$$

That is, person i maximizes his temporal payoff function H^{ii}, by controlling s_a or s_b when he takes role a or b, respectively.

Theorem 1 (Partial-Use Equilibrium). *Suppose (9). The regular pair (s_a^o, s_b^o) is a PUE in Γ^{ii} 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 $s_a \in S_a^i$ and $s_b \in S_b^i$,

$$\rho_{ia} h_a^{ii}(s_a^o, s_b^o) + \rho_{ib} h_b^{ii}(s_a^o, s_b^o) \geq \rho_{ia} h_a^{ii}(s_a, s_b^o) + \rho_{ib} h_b^{ii}(s_a^o, s_b). \tag{15}$$

By this, the *if* part is straightforward. We show the contrapositive of the *only-if* part. Suppose that (s_a^o, s_b^o) is not a NE in g^{ii}. Then, there is some $s_a \in S_a^i$ or $s_b \in S_b^i$ such that $h_a^{ii}(s_a^o, s_b^o) < h_a^{ii}(s_a, s_b^o)$ or $h_b^{ii}(s_a^o, s_b^o) < h_b^{ii}(s_a^o, s_b)$, respectively. Consider the case $h_a^{ii}(s_a^o, s_b^o) < h_a^{ii}(s_a, s_b^o)$. Suppose $\rho_{ia} \neq 0$. In this case, (15) does not hold if we plug (s_a, s_b^o) to the first term of the right-hand side but (s_a^o, s_b^o) to the second term. Suppose $\rho_{ia} = 0$. Then, $\theta_a^i(s_a^o, s_b^o) = h_a^{ii}(s_a^o, s_b^o) < h_a^{ii}(s_a, s_b^o) = \theta_a^i(s_a, s_b^o)$, which violates (9). □

A PUE in the temporal d-understanding Γ^{ii} is reduced into a NE in the static d-understanding g^{ii}; the conjectural postulate $(*)$ is crucial for this reduction. As

far as person i ignores Γ^{ij} but adopts the conjectural postulate $(*)$, the resulting outcome is a NE in the static g^{ii}.

It would be logically possible to have the domain D_i to be the active-passive domain D_i^{AP} given by (5). In this case,

$$h_r^{ii}(s_r; s^o_{-r}) = h_r(s_r; s^o_{-r}) \text{ for all } s_r \in S_r, r = a, b,$$

and it follows from Theorem 1 that every PUE (s^o_a, s^o_b) is a Nash equilibrium in the objective game G. Nevertheless, if role-switching is reciprocal such as having the active-passive domain D_i^{AP}, it would be more natural to expect the use of the *tp*-understanding Γ^{ij}. That is, our main intention for C0 is to handle the non-reciprocal case where $\rho_{ir} = 1$ for $r = a$ or b. In this case, Theorem 1 describes only the side of person i. It would be natural to apply the theorem to both persons, which will be done in Section 5.3.

5. Intrapersonal Coordination Equilibria

Now, our concern is the full use of the temporal i.d.view $\Gamma^i = (\Gamma^{ii}, \Gamma^{ij})$. Since Γ^{ij} allows person i to have some prediction about person j's behavior, we can have an alternative conjectural postulate, to $(*)$, for his behavioral revision, which is given in Section 5.1. This leads to an equilibrium concept called *ICE*. In the remaining, we study its behavior, and obtain a result called the *utilitarian theorem*.

5.1. *Intrapersonal Coordination Equilibria*

For the full use case, we consider the following conjectural postulate: If person i deviates from s^o_r to s_r, then

$(**)$: person j takes also the same deviation s_r at role r.

We call $(**)$ the *role model postulate.*

Conjecture $(**)$ occurs in the mind of person i having his i.d.view $\Gamma^i = (\Gamma^{ii}, \Gamma^{ij})$. No direct communications between i and j are involved in $(**)$. Person i thinks, in his mind, about person j's reaction to i's deviation from the regular actions (s^o_a, s^o_b). We divide this postulate into two sub-postulates:

$(**1)$ (**Thinking in the Same Manner)**: Γ^{ij} supports j's deviation s_r at role r just as Γ^{ii} does for i's deviation s_r at role r.

$(**2)$ (**Initiative Role)**: Person i takes his intentional deviation s_r so as to indicate it to person j to follow.

Fig.4 describes a situation where person 1 deviates from s^o_a to s_a at role a, indicating person 2 to follow this deviation s_a when 2 takes role a,. Let us focus on deviation s_a at role a : the other case for role b is symmetric. The first sub-postulate is formulated as two inequalities:

$$H^{ii}((s^o_a, s^o_b)_a, (s^o_a, s^o_b)_b) < H^{ii}((s_a, s^o_b)_a, (s_a, s^o_b)_b); \tag{16}$$

$$H^{ij}((s_a^o, s_b^o)_a, (s_a^o, s_b^o)_b) < H^{ij}((s_a, s_b^o)_a, (s_a, s_b^o)_b). \tag{17}$$

The d-understanding Γ^{ii} tells person i that his deviation s_a is beneficial, assuming that person j follows the same deviation when j takes role a (i takes b), and the tp-understanding Γ^{ij} tells person i that the symmetric statement holds for person j. Therefore, person i has reason to believe that person j is thinking in the same manner.

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

Fig. 4.

Even though person i believes that a deviation s_a from s_a^o is beneficial for both persons, he needs to assume that when he actually takes this deviation s_a, person j would follow it. In other words, once i takes the initiative to intentionally deviate to s_a at role r, he presumes that j will follow his initiative when j is next at role r. This is postulated in the initiative role postulate $(**2)$.

In Fig.4, person 1 at role a deviates from s_a^o to s_a based on his evaluations (16) and (17). This new situation is expressed as the second left state in Fig.4. It can be thought in the mind of person 1 : Person 2 will observe this deviation, and when he is assigned to role a, he would follow this mutually beneficial deviation s_a, which is the third state in Fig.4. Thus, person 1 can take an *initiative* to deviate from (s_a^o, s_b^o) to s_a : This initiative story motivates the term "role model".

Now, we say that $s_r \in S_r^i$ is a *coordinately improving deviation* (*c-improving deviation*) from the regular actions (s_a^o, s_b^o) iff (16) and (17) hold with the replacement of s_r^o by s_r. For our equilibrium concept, we allow the weaker form of (16) and (17), i.e., the weak inequalities in both (16) and (17) with a strict inequality for at least one person. When these hold, we call $s_r \in S_r^i$ a *weakly c-improving deviation.* By allowing for weak c-improving deviations, we avoid some difficulties arising when H^{ii} or H^{ij} takes constant values.

We now state the definition of our equilibrium concept in the full use of an i.d.view $\Gamma^i = (\Gamma^{ii}, \Gamma^{ij})$:

Definition 3 (ICE). *The regular pair* (s_a^o, s_b^o) *is* an intrapersonal coordination equilibrium *(*ICE*)* in $\Gamma^i = (\Gamma^{ii}, \Gamma^{ij})$ *iff there is no weak* c*-improving unilateral deviation from* (s_a^o, s_b^o).

The term "intrapersonal" is motivated by the fact that possible deviations are all considered by person i's mind, and "coordination" is by the role model postulate.

Now, consider the behavior of an ICE. Our first formal result is about the existence of an ICE. We consider a condition for a strategy pair (s_a^o, s_b^o):

$$h_a(s_a^o, s_b^o) + h_b(s_a^o, s_b^o) \geq h_a(s_r; s_{-r}^o) + h_b(s_r; s_{-r}^o) \text{ for all } s_r \in S_r \text{ and } r = a, b. \tag{18}$$

We call a pair (s_a^o, s_b^o) satisfying (18) a *unilateral utilitarian point (UUP)*. The unweighted payoff-sum takes a maximum at (s_a^o, s_b^o) over the unilateral deviations from (s_a^o, s_b^o). The global maximum of the sum always exists and satisfies (18). Using this condition, we have the following existence theorem of an ICE.

Theorem 2 (Existence). *Let $\Gamma^i = (\Gamma^{ii}, \Gamma^{ij})$ be the i.d.view from a memory kit $\kappa_i = \langle (s_a^o, s_b^o), (D_{ia}, D_{ib}), (h_{ia}, h_{ib}), (\rho_{ia}, \rho_{ib}) \rangle$.*

(1) Let $0 < \rho_{ia} < 1$. Suppose (8). If (s_a^o, s_b^o) is a UUP, then it is an ICE in $\Gamma^i = (\Gamma^{ii}, \Gamma^{ij})$.

(2) Let $\rho_{ir} = 1$ for $r = a$ or b. Suppose (9). Then, (s_a^o, s_b^o) is an ICE in the i.d.view Γ^i if and only if it is a Nash equilibrium in person i's d*-understanding g^{ii}.*

Proof. (1): Since $0 < \rho_{ia} < 1$, we have $D_{ia} \neq \emptyset$ and $D_{ib} \neq \emptyset$ by (2). By Lemma 1, we have $(s_a^o, s_b^o) \in D_{ia} \cap D_{ib}$.

Suppose that (s_a^o, s_b^o) is not an ICE. Then, there is some weak c-improving unilateral deviation, say s_a, from (s_a^o, s_b^o). The other case is symmetric. Now, we have:

$$\begin{aligned} H^{ii}((s_a^o, s_b^o)_a, (s_a^o, s_b^o)_b) &\leq H^{ii}((s_a, s_b^o)_a, (s_a, s_b^o)_b); \\ H^{ij}((s_a^o, s_b^o)_a, (s_a^o, s_b^o)_b) &\leq H^{ij}((s_a, s_b^o)_a, (s_a, s_b^o)_b), \end{aligned} \tag{19}$$

where at least one holds with a strict inequality. Since $(s_a^o, s_b^o) \in D_{ia} \cap D_{ib}$, we have $h_r^{ii}(s_a^o, s_b^o) = h_r^{ij}(s_a^o, s_b^o) = h_r(s_a^o, s_b^o)$ for $r = a, b$. If $(s_a, s_b^o) \notin D_{ia} \cap D_{ib}$, then $h_r^{ij}(s_a, s_b^o) = \theta_r^i(s_a, s_b^o)$ for $r = a, b$, so the second inequality in (19) becomes $(1 - \rho_{ia})h_a(s_a^o, s_b^o) + \rho_{ia}h_b(s_a^o, s_b^o) \leq (1 - \rho_{ia})\theta_a^i(s_a, s_b^o) + \rho_{ia}\theta_b^i(s_a, s_b^o)$, which is impossible by (8). Hence, $(s_a, s_b^o) \in D_{ia} \cap D_{ib}$. Thus, the two inequalities of (19) are expressed as:

$$\begin{aligned} \rho_{ia}h_a(s_a^o, s_b^o) + (1 - \rho_{ia})h_b(s_a^o, s_b^o) &\leq \rho_{ia}h_a(s_a, s_b^o) + (1 - \rho_{ia})h_b(s_a, s_b^o) \\ (1 - \rho_{ia})h_a(s_a^o, s_b^o) + \rho_{ia}h_b(s_a^o, s_b^o) &\leq (1 - \rho_{ia})h_a(s_a, s_b^o) + \rho_{ia}h_b(s_a, s_b^o), \end{aligned}$$

where at least one holds with a strict inequality. Summing up these inequalities, we have $h_a(s_a^o, s_b^o) + h_b(s_a^o, s_b^o) < h_a(s_a, s_b^o) + h_b(s_a, s_b^o)$, which contradicts (s_a^o, s_b^o) being a UUP.

(2): Let $\rho_{ia} = 1$; the other case is symmetric. Then, for all $(s_a, s_b) \in S_a^i \times S_b^i$,

$$H^{ii}((s_a, s_b)_a, (s_a, s_b)_b) = h_a^{ii}(s_a, s_b) \text{ and } H^{ij}((s_a, s_b)_a, (s_a, s_b)_b) = h_b^{ij}(s_a, s_b). \tag{20}$$

Since, by (7) and (9), $h_b^{ij}(s_a, s_b)$ is constant over $\mathrm{Proj}(S_a^i \times S_b^i)$, it follows that (s_a^o, s_b^o) is an ICE if and only if $h_a^{ii}(s_a^o, s_b^o) \geq h_a^{ii}(s_a, s_b^o)$ for all $s_a \in S_a^i$. The latter part is equivalent to that (s_a^o, s_b^o) is a Nash equilibrium in g^{ii}, because $h_b^{ii}(s_a^o, s_b)$ is also constant for $s_b \in S_b^i$ by (6) and (9). □

By Theorem 2, we have the general existence result. In the case of $0 < \rho_{ia} < 1$, the existence of a global maximum of the unweighted sum of payoffs guarantees the

existence of an ICE. In the case of $\rho_{ir} = 1$, if (s_a^o, s_b^o) satisfies the simple payoff maximization over S_r, i.e., $h_r(s_r^o; s_{-r}^o) \geq h_r(s_r; s_{-r}^o)$ for all $s_r \in S_r$, then (s_a^o, s_b^o) is an NE in g^{ii}. For any s_{-r}^o, we can find such a s_r^o. Thus, we have also the existence of an ICE for $\rho_{ir} = 1$.

Theorem 2.(1) allows S_a^i and S_b^i to be proper subsets of S_a and S_b. We can relax (18) for (1) so that it is a maximization condition relative to proper subsets of S_a and S_b. However, this includes the extreme case where $D_{ia} = D_{ib} = \{((s_a^o, s_b^o)\}$, which is not really our target. In Section 5.2, we will restrict our target to the case where $S_a^i = S_a$ and $S_b^i = S_b$ for presentational simplicity.

5.2. *Utilitarian Theorem*

Theorem 2.(1) states that a UUP is an ICE for any (non-extreme) frequency weights (ρ_{ia}, ρ_{ib}). We expect the converse, though, rigorously speaking, we need some assumptions for it.

First, we restrict our attention to internally reciprocal domains, i.e., those satisfying (3). We also assume, for simplicity that,

$$S_a^i = S_a \text{ and } S_b^i = S_b. \tag{21}$$

It follows from this, Lemma 3 and (2) that all the payoffs from (s_a^o, s_b) and (s_a, s_b^o) for $s_a \in S_a$ and $s_b \in S_b$ are experienced. Hence, we do not need to refer to condition (8) or (9).

We assume the following genericity condition on the base game G: for all $s_r' \in S_r$ with $s_r' \neq s_r^o$ $(r = a, b)$,

$$h_a(s_a^o, s_b^o) + h_b(s_a^o, s_b^o) \neq h_a(s_r'; s_{-r}^o) + h_b(s_r'; s_{-r}^o). \tag{22}$$

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

Let $\Gamma^i = (\Gamma^{ii}, \Gamma^{ij})$ be the i.d.view from a memory kit κ_i satisfying (3)(internal reciprocity). Then, $(s_a^o, s_b^o) \in D_{ia} \cup D_{ib}$, and so by (2), $0 < \rho_{ia} < 1$. Let us assume $0 < \hat{\rho}_{ia} < 1$ and $\hat{\rho}_{ib} = 1 - \hat{\rho}_{ia}$. We obtain $\Gamma^{ii}(\hat{\rho}_{ia}, \hat{\rho}_{ib})$ and $\Gamma^{ij}(\hat{\rho}_{ia}, \hat{\rho}_{ib})$ from Γ^{ii} and Γ^{ij} by substituting $(\hat{\rho}_{ia}, \hat{\rho}_{ib})$ for (ρ_{ia}, ρ_{ib}) in Γ^{ii} and Γ^{ij}, and we write $\Gamma^i(\hat{\rho}_{ia}, \hat{\rho}_{ib}) = (\Gamma^{ii}(\hat{\rho}_{ia}, \hat{\rho}_{ib}), \Gamma^{ij}(\hat{\rho}_{ia}, \hat{\rho}_{ib}))$.

Theorem 3 (Utilitarian Theorem). *Let $\Gamma^i = (\Gamma^{ii}, \Gamma^{ij})$ be the i.d.view from a memory kit κ_i satisfying (3) and (21).*

***(1)**: (s_a^o, s_b^o) is an ICE in $\Gamma^i(\frac{1}{2}, \frac{1}{2})$ if and only if it is a UUP.*

***(2)**: Suppose (22) for the game G. Then, (s_a^o, s_b^o) is a UUP if and only if there are α, β with $0 < \alpha < 1/2 < \beta < 1$ such that (s_a^o, s_b^o) is an ICE of $\Gamma^i(\hat{\rho}_{ia}, \hat{\rho}_{ib})$ for all $\hat{\rho}_{ia} \in [\alpha, \beta]$.*

Proof. (1): Since $\text{Proj}(D_{ia}) = \text{Proj}(D_{ib})$ by (3), we have $(s_r; s_{-r}^o) \in D_{ia} \cap D_{ib}$ for all $s_r \in S_r^i, r = a, b$. Therefore, the payoffs h_r^{ii} and h_r^{ij} coincide with h_r for

the whole domain of unilateral deviations. This implies that (s_a^o, s_b^o) is an ICE in $\Gamma^i(\frac{1}{2}, \frac{1}{2})$ if and only if $\frac{1}{2}h_a(s_a^o, s_b^o) + \frac{1}{2}h_b(s_a^o, s_b^o) \geq \frac{1}{2}h_a(s_r; s_{-r}^o) + \frac{1}{2}h_b(s_r; s_{-r}^o)$ for all $s_r \in S_r, r = a, b$. This is equivalent to the condition that (s_a^o, s_b^o) is a UUP.

(2): Let (s_a^o, s_b^o) be a UUP. Then, by (22), we have $\frac{1}{2}h_a(s_a^o, s_b^o) + \frac{1}{2}h_b(s_a^o, s_b^o) > \frac{1}{2}h_a(s_r; s_{-r}^o) + \frac{1}{2}h_b(s_r; s_{-r}^o)$ for all $s_r \in S_r - \{s_r^o\}, r = a, b$. Hence, there are some α, β with $0 < \alpha < 1/2 < \beta < 1$ such that for any $\hat{\rho}_{ia} \in [\alpha, \beta]$,

$$\begin{aligned} &\hat{\rho}_{ia}h_a(s_a^o, s_b^o) + (1 - \hat{\rho}_{ia})h_b(s_a^o, s_b^o) \geq \hat{\rho}_{ia}h_a(s_r; s_{-r}^o) + (1 - \hat{\rho}_{ia})h_b(s_r; s_{-r}^o); \quad (23) \\ &(1 - \hat{\rho}_{ia})h_a(s_a^o, s_b^o) + \hat{\rho}_{ia}h_b(s_a^o, s_b^o) \geq (1 - \hat{\rho}_{ia})h_a(s_r; s_{-r}^o) + \hat{\rho}_{ia}h_b(s_r; s_{-r}^o). \end{aligned}$$

for all $s_r \in S_r, r = a, b$. These imply that (s_a^o, s_b^o) is an ICE of $\Gamma^i(\hat{\rho}_{ia}, \hat{\rho}_{ib})$ for any $\hat{\rho}_{ia} \in [\alpha, \beta]$.

For the converse, suppose that (s_a^o, s_b^o) is not a UUP, and let α, β satisfy $0 < \alpha < 1/2 < \beta < 1$. Then, by (1) of this theorem, (s_a^o, s_b^o) is not an ICE of $\Gamma^i(\frac{1}{2}, \frac{1}{2})$□

Theorem 3 asserts the equivalence between the ICE and UUP under the reciprocal domains condition (3) in some neighborhood of $\hat{\rho}_{ia} = \frac{1}{2}$. Even when the frequency weights $(\hat{\rho}_{ia}, \hat{\rho}_{ib})$ are skewed, the UUP is determined independent of this skewedness. Although person i detects some skewedness, the resulting outcome is free from it. Thus, we have an unweighted utilitarian (up to unilateral domains) outcome for an ICE, which motivates the title of the Theorem 3.

It may be questioned how large the size of the interval $[\alpha, \beta]$ in Theorem 3.(2) is, and also what would happen outside the interval. We consider these problems in the PD and SH games given in Section 1. For this consideration, let us denote the infimum and supremum of such α's and β's in Theorem 3.(2) by α_0 and β_0, respectively. These can be calculated from the inequality system (23).

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

Prisoner's Dilemma: Consider Table 1. By Theorem 2.(1), the regular pair $(\mathbf{s}_{a1}, \mathbf{s}_{b1})$ is an ICE point for any $\hat{\rho}_{ia} \in (0, 1)$. Since this is the unique UUP, it follows by Theorem 3.(2) that there is some interval $[\alpha, \beta]$ such that for $\hat{\rho}_{ia} \in [\alpha, \beta]$, $(\mathbf{s}_{a1}, \mathbf{s}_{b1})$ is the unique ICE of $\Gamma^i(\hat{\rho}_{ia}, \hat{\rho}_{ib})$. 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 $(\mathbf{s}_{a1}, \mathbf{s}_{b2})$, $(\mathbf{s}_{a2}, \mathbf{s}_{b2})$ and $(\mathbf{s}_{a1}, \mathbf{s}_{b2})$ appear as ICE points.

Stag Hunt: The SH1 game of Table 2 has the unique UUP, $(\mathbf{s}_{a1}, \mathbf{s}_{b1})$, which is an ICE for all $\hat{\rho}_{ia} \in (0, 1)$. 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 NE (s_{a2}, s_{b2}) appears as an ICE point.

In the SH2 game of Table 3, we have two UUP's, $(\mathbf{s}_{a1}, \mathbf{s}_{b1})$ and $(\mathbf{s}_{a2}, \mathbf{s}_{b2})$. Both are ICE points for any frequency $\hat{\rho}_{ia} \in (0, 1)$. We have no other candidates for an ICE point.

5.3. *Mutual ICE and Mutual PUE*

We have considered the behavioral consequences derived from one person's perspective. These arguments can be applied to both persons. Here, we give a small discussion on those applications. We say that $\Gamma^1 = (\Gamma^{11}, \Gamma^{12})$ and $\Gamma^2 = (\Gamma^{22}, \Gamma^{21})$ are *mutually coherent* iff $\rho_{1a} = 1 - \rho_{2a}$. That is, the perceived frequency weights (ρ_{1a}, ρ_{1b}) and (ρ_{2a}, ρ_{2b}) of persons 1 and 2 are consistent with each other. We say that the pair (s_a^o, s_b^o) of regular actions is a *mutual ICE* iff it is an ICE in Γ^i for $i = 1, 2$.

We have the following corollary from Theorem 2.

Corollary 1 (Existence of a Mutual ICE). *Let Γ^1, Γ^2 be the mutual coherent i.d.views.*

***(1)**: Let $0 < \rho_{ia} < 1$ for $i = 1, 2$. Suppose (8) and assume that (s_a^o, s_b^o) is a UUP. Then, (s_a^o, s_b^o) is a mutual ICE.*

***(2)**: Let $\rho_{1a} = \rho_{2b} = 1$. Suppose (9) and assume (21) for each Γ^{ii}. Then, (s_a^o, s_b^o) is a mutual ICE if and only if it is a Nash equilibrium in the base game G.*

We have a parallel result in the partial use case C0. We say that the pair (s_a^o, s_b^o) of regular actions is a *mutual PUE* iff it is a PUE in Γ^{ii} for $i = 1, 2$. Then we have the following corollary from Theorem 1.

Corollary 2 (Mutual PUE for Non-reciprocal Cases). *Let Γ^1, Γ^2 be the mutually coherent i.d.views with $\rho_{1a} = 1$. Suppose (9) and assume (21) for each Γ^{ii}. Then, (s_a^o, s_b^o) is a mutual PUE if and only if it is a Nash equilibrium in the game G.*

6. Extensions and Further Applications

We have already seen some applications of our theory to the PD and SH games in Section 5. In this section, first we mention some experimental results on PD and SH 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.* [2013] undertook experiments for the cases of no role-switching and full alternating role-switching for some PD and SH 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 trials/errors 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.

Postulates BH1, BH2 are supported in that the trajectories of actions taken by many pairs of subjects showed some convergences while keeping some deviations. Comparing the trajectories with the reported payoff values for the questionnaires, we find also that postulates EP1, EP2 are also supported and sharpened in that more experienced payoffs are more accurately reported, but some reported payoffs are incorrect. The predictions given in Section 5 are supported by the combinations of revealed behavior and reported (even incorrect) payoffs in the experiments. While incorrect reporting of payoffs is consistent with postulates EP1 and EP2, theoretically speaking, the theory given in this paper does not treat incorrect understanding of payoffs. A study of inductively derived views including incorrect payoffs remains open.

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 α_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.5.

The game has a unique backward induction solution: $((99,1),(Y,Y,Y))$. This is quite incompatible with experimental results (cf., Camerer [2003]), 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

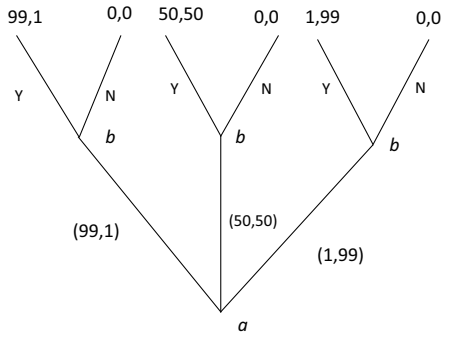

Fig. 5. Ultimatum Game

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

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

The inequality follows from the strict concavity of u. In this game, an ICE is given as $((50,50),(\alpha_1,Y,\alpha_3))$, where α_1, α_3 may be Y or N.

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 Kaneko and Kline [2008a], Kaneko and Kline [2008b] in order to possibly reduce the set of ICE's.

In the extensive form game of Fig.5, person j at role b will be able to observe the deviation of i at role a before j makes his move within the same round of play of the game. This differs from the strategic game case, where the observation by j comes only after both players have moved as described by assumption Ob. This difference allows for the possibility of joint deviations occurring within the same round of play. 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 such joint deviations for the ICE, $((50,50),(\alpha_1,Y,\alpha_3))$ are the only 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 constrained maximization of additional objective functions (cf., Camerer [2003]). 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 3. It has a similarity with the "moral sentiments" due to Smith [1759] 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 thoughts through role-switching and transpersonal projection.

7. External and Internal Reciprocities

Internal reciprocity (3), which was used as a key condition for Theorems 2 and 3, represents reciprocity of the domains of accumulation (D_{ia}, D_{ib}) within one person i. In this section we show that internal reciprocity can be motivated and derived from entirely external comparisons between the domains of persons i and 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 fact, we could impose the following condition on domains of accumulation: For all $s_r \in S_r$, $r = a, b$ and $i = 1, 2$,

$$(s_r; s^o_{-r}) \in D_{j(-r)} \text{ implies } (s_r; s^o_{-r}) \in D_{ir}. \tag{24}$$

That is, if j at role $-r$ keeps a passive experience $(s_r; s^o_{-r})$, then i keeps the same pair as an active experience. This one directional implication means that a person is more sensitive to being active than passive. Condition (24) 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 (4) satisfy (24).

As time passes, each person may have learned also passive experiences. Eventually, the converse of (24) could hold: For all $s_r \in S_r$, $r = a, b$ and $i = 1, 2$,

$$(s_r; s^o_{-r}) \in D_{j(-r)} \text{ if and only if } (s_r; s^o_{-r}) \in D_{ir}. \tag{25}$$

The non-reciprocal active domains D^N_1 and D^N_2 fail to satisfy (25), but this does not yet imply the internal reciprocity of (3). Condition (25) requires the two persons to have the same sensitivities to experiences, but allows them to have different trial deviations. We can have another interpersonal condition requiring the same trial deviations: for all $s_r \in S_r$, $r = a, b$ and $i = 1, 2$,

$$(s_r; s^o_{-r}) \in D_{jr} \text{ if and only if } (s_r; s^o_{-r}) \in D_{ir}. \tag{26}$$

Together, (25) and (26) require the same sensitivities to experiences and the same trial deviations. Both are external relationships, but they are enough to guarantee the internal reciprocity of (3), and also a form of external reciprocity. We say that $D_1 = (D_{1a}, D_{1b})$ and $D_2 = (D_2, D_{2b})$ are *externally reciprocal* iff

$$\text{Proj}(D_{1r}) = \text{Proj}(D_{2r}) \text{ for } r = a, b. \tag{27}$$

We now have two types of reciprocities: internal reciprocity (3) within a person, and external reciprocity (27) across people. When these two types of reciprocity are

taken together, they are shown to be equivalent to the two external conditions (25) and (26) on the sensitivities and trial deviations of the two persons. This equivalence is stated in the following theorem.

Theorem 4 (Internal-External Reciprocity). *Let $D_1 = (D_{1a}, D_{1b})$ and $D_2 = (D_{2a}, D_{2b})$ be the domains of accumulation of persons 1 and 2. Conditions (25) and (26) hold if and only if (3) and (27) hold.*

Proof. When (3) and (27) hold, the four sets, $\mathrm{Proj}(D_{ir})$, $i = 1, 2$ and $r = a, b$ coincide. Hence, the *if*-part is straightforward. We prove the *only-if* part. Suppose (25) and (26) for D_1 and D_2.

Consider (3). Let $(s_a, s_b) \in \mathrm{Proj}(D_{1a})$, i.e., $(s_a, s_b) = (s_a, s_b^o)$ or $(s_a, s_b) = (s_a^o, s_b)$. Let $(s_a, s_b) = (s_a, s_b^o)$. Then, $(s_a, s_b^o) \in \mathrm{Proj}(D_{2a})$ by (26), and so by (25), we have $(s_a, s_b^o) \in \mathrm{Proj}(D_{1b})$. The case of $(s_a, s_b) = (s_a^o, s_b)$ is shown in a similar way first applying (25), and then (26). Thus, we have $\mathrm{Proj}(D_{ia}) \subseteq \mathrm{Proj}(D_{ib})$. The converse is obtained by a symmetric argument.

Consider (27). Let $(s_a, s_b) \in \mathrm{Proj}(D_{1a})$, i.e., $(s_a, s_b) = (s_a, s_b^o)$ or $(s_a, s_b) = (s_a, s_b^o)$. First, let $(s_a, s_b) = (s_a^o, s_b)$. By (26), we have $(s_a, s_b^o) \in \mathrm{Proj}(D_{2a})$. Next, let $(s_a, s_b) = (s_a, s_b^o)$. By (26), we also have $(s_a^o, s_b) \in \mathrm{Proj}(D_{2a})$. We have shown that $\mathrm{Proj}(D_{1a}) \subseteq \mathrm{Proj}(D_{2a})$. The converse is obtained by a symmetric argument□

We interpreted frequency weights as subjective perception. To study the relationships between these and internal/external reciprocities on D_i's, we should refer to objective frequency weights. In the experimental study of Takeuchi *et al.* [2013], 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 frequency $\frac{1}{3}$ for role a. Then, person i experiences role b twice more than role a; he needs 2 times longer than for a to have the same number of experiences of b. By Postulate EP1, he may forget previous experiences from role a. At a point of time, the domain D_{1a} may be much smaller than D_{1b}. In this case, the internal reciprocity of (3) may not be expected.

8. Conclusions

We have introduced the concept of social roles into IGT in order to study an experiential foundation of a person's understanding of his own situation and the thought's of the other. Based on this foundation, we have shown the possibility for the emergence of cooperation, and argued that cooperation is more likely to be achieved when role-switching is more reciprocal. The foundational study and cooperation result have implications to the three important literatures: (a) the argument by

Mead [1934] for role-switching and cooperation, (b) cooperative game theory, and (c) noncooperative game theory from the perspective of *ex ante* decision making. Since our analysis is restricted to the 2-person case, we first give a comment on this restriction before talking about (a), (b) and (c).

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. It would be difficult, from the perspective of finite and bounded cognitive abilities of persons, for a player to experience and 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: Each social situation is not isolated from other social situations in the entire social web as depicted in Fig.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 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 base, etc. Also, once a pitcher understands the perspective of a third baseman, he may extend his understanding to the other infielders.

This argument is quite different from the cooperative game theory literature from von Neumann and Morgenstern [1944]. 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 some literature such as that of "assignment games" initiated by Shapley and Shubik [1971], 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 discussed above.

Bacharach [1999] developed a theory of team reasoning, which may look similar to our ICE concept. This theory starts with given possible team cooperation, and discusses reasonings by individual members in a team. The theory is formulated as a game model with incomplete information. Our emphasis is on the experiential origin and emergence of one's own and the other's understanding of the game situation. Even if we forget our emphasis, the similarity between the equilibrium concept ICE and team reasoning is similar only in that both treat cooperation in a noncooperative manner. To analyze the logical reasoning about the other's reasoning, a more recent development of epistemic logic (cf., Fagin *et al.* [1995], Kaneko [2002]) should be more relevant than Bacharach's approach.

Finally, we should give comments on (c). 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 *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 *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 1), but is that we did not take the perspective of *ex ante* decision making for our cooperative result. To have our cooperation result, we needed the dynamic feature of the frequency weights for role-switching and the average payoffs.

Nonetheless, since our concept of a transpersonal view treats higher-order beliefs, some reader 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 and Zamir [1985] and Brandenburger and Dekel [1993]), and the other is the epistemic logic approach (cf., Fagin *et al.* [1995], Kaneko [2002], and Kaneko and Suzuki [2002]). 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 any further. Instead, it would be helpful to discuss whether the universal-type space approach or the epistemic logic approach is more natural for an explicit treatment of the source of beliefs.

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 and Suzuki [2002] 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. While

the epistemic logic approach shows promise in treating these matters, and some connections are being made, there is still much to be done.

Acknowledgments

The authors are partially supported by Grant-in-Aids for Scientific Research No.21243016 and No.2312002, Ministry of Education, Science and Culture, and Australian Research Council Discovery Grant DP1100103884. The authors thank Burkhard Schipper, Yusuke Narita, Akihiko Matsui for some helpful comments, and also the referee for various constructive suggestions.

References

Akiyama, E., R. Ishikawa, M. Kaneko and J. J. Kline, [2013], Inductive Game Theory: A Simulation Study of Learning a Social Situation, *Game Theory Relaunched*, edited by H. Hanappi, 55-76. InTech - open science, open minds.

Bacharach, M. [1999], Interactive Team Reasoning: A Contribution to the Theory of Cooperation, *Research in Economics* 53, 117-147.

Brandenburger, A., and E. Dekel [1993], Hierarchies of Beliefs and Common Knowledge, *Journal of Economic Theory*, 59, 189-198.

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

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

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

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

Hu, T.-W., [2013], Expected Utility Theory form the Frequentist Perspective, *Economic Theory* 53, 9-25.

Kaneko, M., [2002], Epistemic Logics and their Game Theoretical Applications: Introduction. *Economic Theory* 19, 7-62.

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

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

Kaneko, M., and J. J. Kline, [2013], Partial Memories, Inductively Derived Views, and their Interactions with Behavior, *Economic Theory* 53, 27-59.

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

Kaneko, M., and N.-Y. Suzuki, [2002], Bounded Interpersonal Inferences and Decision Making, *Economic Theory* 19, 63-103.

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

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

Mertens, J., and S. Zamir [1985], Formulation of Bayesian Analysis for Games with Incomplete Information, *International Journal of Game Theory* 14, 1-29.

Osborne, M. J. and A. Rubinstein, [1994], *A Course in Game Theory*, MIT Press, London.

Shapley, L., and M. Shubik, [1971], Assignment Game I: The Core, *International Journal of Game Theory* 1, 111-130.

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

Takeuchi, A., Y. Funaki, M. Kaneko, and J. J. Kline, [2011] An Experimental Study of Behavior and Cognition from the Perspective of Inductive Game Theory, DP.1267, http://www.sk.tsukuba.ac.jp/SSM/libraries/pdf1251/1267.pdf

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