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

Mamoru Kaneko[‡]and J. Jude Kline [§]

26 March 2009

Abstract

Inductive game theory has been developed to explore the origin/emergence of beliefs/knowledge 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 origin/emergence of one's view of the other's beliefs/knowledge about the game situation. We restrict our exploration to a 2-role (strategic) game, which has been recurrently played by two persons with switching the roles of positions (players). By switching roles, each person accumulates experiences of both roles and these experiences become the source of a person's (transpersonal) view about the other person's view. We will show that as the degree of reciprocity increases, cooperation will emerge.

1. Introduction

We will consider the problem of how a person obtains beliefs/knowledge about other persons' thoughts. We look for experiential bases for such beliefs/knowledge. A crucial distinction is made between persons (actors) and social roles (players), which allows a person to switch roles from time to time. This enables a person, based on his experiences, to guess the other person's thinking, and even to obtain a social perspective, which goes beyond an individual perspective. Within this framework, we can go further to discuss the emergence of cooperation.

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

[†] The authors thank Nathan Berg for valuable discussions on the subject of the present paper.

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

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

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

1.1. General Motivations

It is customary in game theory and economics to assume well-formed beliefs/knowledge of a game for each player, which is often implicit and sometimes explicit. The present authors [14], [15] and [16] have developed inductive game theory in order to explore the basic question of where a personal understanding of a game comes from[1]. In those papers, an individual view and its derivation from a player's experiences are discussed from various points of view. Nevertheless, they have focussed an individual perspective of society but did not reach the stage of research on his thoughts about other persons' thoughts. This paper aims to take one step further to explore the origin and emergence of a person's thoughts about other persons' thoughts.

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

When the persons switch social roles reciprocally, a new feature is emerging: Reciprocal relationships provide each person with a rich source for inferring/guessing the beliefs of the other person. When the persons switch roles enough, each has been in the same position and has seen the other person in the corresponding position. This level of reciprocity may give each person "reason to believe" that the other's view is the same as his. This idea is reminiscent of a requirement imposed for "common knowledge" in Lewis [19]. This will be the key for the development of our theory.

Broadly speaking, we may regard our exploration as undertaken along the line of *symbolic interactionism* due to Mead [20] (cf., Collins [5], Chap.7). Each isolated experience is not more than a sequence or a set of symbols. However, by playing roles reciprocally and interactively, the accumulated set of experiences could constitute some meaning. This is analogous also to symbolic logic (cf., Mendelson [21] and Kaneko [12]) in that it starts with primitive symbols without meanings. Formulae consisting of those symbols and their further combinations may eventually generate some meanings. An individual perspective is obtained by combining experiences to obtain some meaningful view. A social perspective is obtained by combining experiences of reciprocal

[1] A seminal form of inductive game theory was given in Kaneko-Matsui [17].

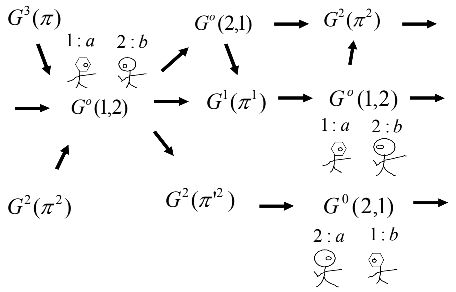

Figure 1.1: Social Web

interactions into an even greater view. In the sociology literature, these problems were discussed without giving a mathematical formulation. Our approach is regarded as a mathematical formulation of symbolic interactionism, and expands its perspective while enabling us to examine it critically.

An example, due to Mead [20], for the distinction between a person and a social role consists of the positions of pitcher, catcher, first base, etc..., in a baseball team. Since we use only a 2-role strategic game for our exploration, it may be better to refer to a 2-role example of a family affair between a wife and a husband: They may divide their housekeeping into the breakfast maker and dinner maker. There are numerous alternative varieties, e.g., raising children versus working at the office, cleaning the house versus gardening, or allocating finances versus generating finances. In such situations, role-switching becomes crucial for understanding the other's perspective.

A target game situation is in a social web like Fig.1.1: Two persons 1 and 2 play the strategic game $G^o(1,2)$ in the north-west in Fig.1.1, where G^o is assumed to be a standard strategic game with two "players", which are roles a and b. In $G^o(1,2)$, persons 1 and 2 take roles a and b, respectively. If they switch roles a and b, the game situation becomes as $G^o(2,1)$ in the south-east. Although we will focus on a particular situation such as G^o, it is a small part of the entire social web for the persons. Each person participates in various other social games such as university administration, a community baseball team, etc. This remark should not be forgotten, and will be discussed in various places in the paper.

If person 1 has experienced two roles a and b from time to time, person 1 could guess person 2's thoughts. This is the source for interpersonal beliefs/knowledge of the structure of the game. Of course, this requires reciprocity of roles played by those two persons. We will explore how such reciprocity is needed for person 1 to fully imagine the other's thoughts. Another extreme case, which should not be ignored, is one where they do not switch roles at all, and as a consequence, person 1 cannot imagine 2's thoughts. Our theory presents some capacity to separate these cases and generates different results based on this separation. One such difference is that cooperation would not be reached without a sufficient level of reciprocity.

It is a salient point of our approach that thinking about the other's thoughts in one's mind might lead to cooperation. This type of idea was discussed and emphasized by Mead [20] and his predecessor, Cooley [6] to argue the pervasiveness of cooperation in human society. This level of optimism was criticized as too naive by later sociologists (see Collins [5], Chap.7). In our theory, cooperation is one possibility, but not necessarily guaranteed. We can discuss when cooperation likely happens and when not.

Another comment is that although we discuss social roles, we do not address the important problem of emergence of "social roles". Instead, we treat social roles as exogenously given, and target the emergence of the other's thoughts and cooperation based on role-switching[2].

In the game theory literature, cooperative behavior has been extensively discussed, but no relationships between cooperative behavior and cognitive assumptions are discussed. Since this will be important to distinguish our new theory from the other extant theories, we will give brief discussions on the treatments of cooperative behavior in the game theory literature in Section 1.2.

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

1.2. Brief Discussions on Cooperative Behavior in the Literature

Cooperative behavior or cooperation has been discussed a lot in the literature of game theory and economics. In contrast to our approach, cooperative behavior there is simply assumed and/or is shown to be derived through noncooperative game theory. In all of these theories, it is, implicitly or explicitly, assumed that the players have beliefs/knowledge on the game structure; they neither aim to nor are able to discuss the emergence of basic beliefs/knowledge. Nevertheless, small summaries of these fields would help the reader to understand what we are going to do and how it relates to the extant theories. Here, we will look only at cooperative game theory, the Nash program, and the repeated game approach.

Cooperative game theory was already extensively discussed in von Neumann- Morgenstern [25] and a lot of branches have been developed. In them, cooperation itself is a very basic postulate, and possible outcomes resulting from cooperative behavior are

[2] This is pointed out by Nathan Berg.

targets to be studied. This theory is incapable of discussing the origin/emergence of cooperation.

The *Nash program*, which was originally suggested by Nash [22], p.295, may appear to resolve this incapability by reducing cooperation into individual activities for cooperation: The possibilities for some players to propose to cooperate with some other players are described as moves in (rules of) an extensive game. In this theory, we may discuss a process for cooperative behavior, an example of which was given in Nash [23]. The Nash program reduces the postulate of cooperation into the rules of a game, but this theory is also incapable of addressing the question of emergence of cooperation.

The *repeated game approach* (cf., Hart [9]) has two similar aspects to our approach in that both treat recurrent situations and discuss cooperation as a possible outcome. Nevertheless, the two approaches have a radical difference in their basic cognitive postulates. The repeated game approach cannot address the question of emergence of beliefs/knowledge. Also, in the repeated game approach, a cooperative outcome is based on threats and the emergence of cooperation/cooperative behavior is not in the scope. In our theory, cooperation is founded on experiential understanding of the roles and behaviors in the game.

The repeated game approach formulates the entire situation as a huge one-shot game, i.e., an infinite extensive game. Then the Nash equilibrium (or its refinement) is adopted for this entire game. The Nash equilibrium is interpreted as describing *ex ante* decision making in the sense that each player makes a decision as well as his prediction about the others' decision before the actual play of the repeated game. This requires each player to be fully cognizant of the entire game structure[3]. For this reason, the repeated game approach cannot address the basic cognitive question of where beliefs/knowledge about the game structure and others' beliefs comes from for a player. In this respect, the Nash program is in the same position; the Nash program also requires full cognizance.

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

Fig.1.2 illustrates what we mean by "emergence of cooperation". We will formulate the concept of an *intrapersonal coordination equilibrium.* It is the same as the non-cooperative Nash equilibrium when the situation is non-reciprocal. When the situation is more reciprocal, some cooperative outcome is emerging. Hence, the degree of reciprocity

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

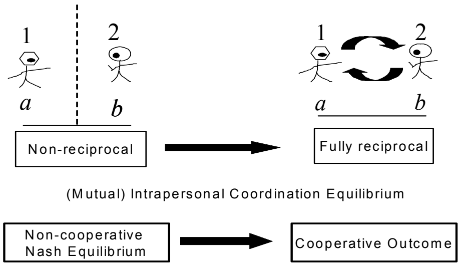

Figure 1.2: Emergence of Cooperation

is a key for the cooperation.

We will connect our cooperation result to some behavioral game theory literature. Behavioral/experimental game theory has reported many experiments to support the pervasiveness of cooperative behavior. One observation in the repeated situation of the prisoner's dilemma is that cooperative outcomes emerge after some repetition of the game (cf., Cooper-DeJong-Forsyth-Ross [4]) Another is the experimental study of the ultimatum game and dictator game, which shows that people do cooperate, even though the standard game theoretical argument (subgame perfection) does not predict cooperation at all (cf., Güth-Schmittberger-Schwarze [7], Kahneman-Knetsch-Thaler [11], and also Camerer [3] for a more recent survey). In Section 7, we will examine implications of our theory of cooperation to the literature of those behavioral studies, specifically, looking at the ultimatum game and dictator game.

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

Kaneko-Kline [14], [15] and [16] chose a general environment corresponding to an extensive game and already met a lot of basic problems in the consideration of experiences and their generations. Also, various basic notions in the extant game theory such as "information", "memory", and moreover, "extensive game" itself needed to be redefined. In the consideration of induction, they met also partiality, indeterminacy, falsity, etc. in an inductively derived view on the social structure.

As stated above, we confine ourselves here to a 2-role strategic game to avoid the difficulties mentioned above. Nevertheless, since we now include the other's thoughts, our theory becomes fairly complex, and also, we need various subtle definitions. Thus, it would be better to mention the basic postulates for one's thinking about the other's and for the emergence of cooperation.

First, we make the *basic postulate* that a person cannot directly look into the other's mind. Instead, we *postulate* that person i infers/guesses from his own experiences what person j may know about the situation. Transpersonal projection of one's experiences onto the other is considered based on experiences of different roles. Thus, our theory is experiential and follows the tradition from Mead [20].

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

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

2. Two-Person Strategic Game with Social Roles

2.1. 2-Role Strategic Game and Role Assignments

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

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

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

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

Example 2.1: In the game $G(1,2)$ of Table 2.1, persons 1 and 2 are assigned to roles a and b. The game $G(2,1)$ has the same structure, but the role-assignments are reversed. The game G of Table2.1 is symmetric with respect to roles a and b. The game G' of Table 2.2 obtained from Table 2.1 by multiplying the payoffs of role b by 2 is a nonsymmetric example. In our theory, this multiplication (more generally, an affine transformation) of payoffs may matter to behavioral results, which will be clear in Section 6.

Table 2.1; $G(1,2)$

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

Table 2.2; G'

$a\backslash b$	$\mathbf{s}_{b1}$	$\mathbf{s}_{b2}$	$\mathbf{s}_{b3}$
$\mathbf{s}_{a1}$	$(3,6)$	$(10,4)$	$(3,2)$
$\mathbf{s}_{a2}$	$(2,20)$	$(4,8)$	$(5,10)$
$\mathbf{s}_{a3}$	$(1,6)$	$(5,10)$	$(4,8)$

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

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

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

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

These postulates are asymmetric in that person i can observe both actions taken by him and the other, but can observe only his own payoff. This asymmetry will be important in Section 3. With respect to the treatment of payoffs, we should emphasize the distinction between *having* a payoff function and *knowing* it. Here, we assume that each person recognizes each payoff value $h_r(s_a, s_b)$ only when he experiences it but does not know the function h_r itself. Only after he has accumulated enough memories of experiences, he may come to know some part of the payoff function.

2.2. Accumulated Memories

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

κ1: the pair (s^o_a, s^o_b) of *regular actions*;

κ2: the *accumulated domain of experiences* $D_i = (D_{ia}, D_{ib})$ consisting of experiences of action pairs from taking roles a and b, respectively;

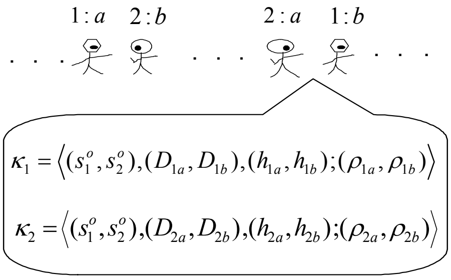
1 : a
2 : b
2 : a
1 : b
$\kappa_1 = \langle (s_1^o, s_2^o), (D_{1a}, D_{1b}), (h_{1a}, h_{1b}); (\rho_{1a}, \rho_{1b}) \rangle$
$\kappa_2 = \langle (s_1^o, s_2^o), (D_{2a}, D_{2b}), (h_{2a}, h_{2b}); (\rho_{2a}, \rho_{2b}) \rangle$

Figure 2.1: Memory Kits

$\kappa3$: person i's *observed payoff functions* (h_{ia}, h_{ib}) over D_i;

$\kappa4$: person i's vector (ρ_{ia}, ρ_{ib}) of *frequency weights* for roles a and b.

Person i has obtained these components by playing game G with possibly different roles from time to time. Component $\kappa1$ means that the persons play regularly the actions s_a^o and s_b^o when they are assigned to roles a and b. Component $\kappa2$ states that person i has other experiences in addition to the regular actions. Occasionally, each person i deviates from s_r^o to some other actions s_r, and some (or all) actions experienced are remaining in his mind, which form the sets D_{ia} and D_{ib}. The third components, (h_{ia}, h_{ib}), in $\kappa3$ are the observed (perceived) payoff functions over (D_{ia}, D_{ib}), which are mathematically defined presently. The last component (ρ_{ia}, ρ_{ib}) in $\kappa4$ means that person i evaluates how frequently he has been assigned to roles a and b. Accurate weights are not really our intention[4], but here we assume that it is a single vector for each i.

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

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

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

$$\text{if } (s_r; s_{-r}) \in D_{ir}, \text{ then } (s_r; s_{-r}^o) \in D_{ir}; \tag{2.3}$$

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

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

Condition (2.2) states that the domains of accumulation include the regular actions (s_a^o, s_b^o). It is the intent that (s_a^o, s_b^o) has been played in $G(1,2)$ and $G(2,1)$ as the regular

[4] See Hu [10] for the concept of frequency and the frequentist interpretation of expected utility theory.

actions, while person i has made some trial deviations from (s_a^o, s_b^o) and accumulated his experiences in D_{ia} and D_{ib}. We allow D_{ia} or D_{ib} to be empty, though the union $D_{ia} \cup D_{ib}$ is nonempty by (2.2). If $D_{ia} = \emptyset$, then person i has never experienced role a at least in his memory.

Condition (2.3) that if ever some pair $(s_r; s_{-r})$ is accumulated in D_{ir}, then the pair coming from the unilateral trial s_r from the regular action s_r^o is also accumulated. It expresses the idea that the domain of accumulation is generated by unilateral trials from the regular action. This will be briefly discussed in Section 2.3.

Condition (2.4) states that person i knows a functional relationship between each pair $(s_a, s_b) \in D_{ir}$ and the payoff value from it when he takes role r. To avoid confusions with the objective payoff function h_r, we define the function $h_{ir} : D_{ir} \to \mathbf{R}$. Thus, this is the experienced payoff function of person i when he takes role r. Mathematically, h_{ir} is simply the restriction of h_r to D_{ir}. Finally, (2.5) states that (ρ_{ia}, ρ_{ib}) is a vector of frequency weights, and does not exclude the possibility of either ρ_{ia} or ρ_{ib} being 0, which is the non-reciprocal case.

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

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

$$D_{ia} = D_{ib}. \tag{2.6}$$

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

In fact, (2.6) is stronger than what we will target in this paper. For the weaker version, first we define the set $\text{Proj}(T) := \{(s_a, s_b) \in T : s_a = s_a^o \text{ or } s_b = s_b^o\}$. Then, (2.6) is weakened to

$$\text{Proj}(D_{ia}) = \text{ Proj}(D_{ib}), \tag{2.7}$$

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

We should bear in mind that since the experiences in $D_{ia} \cup D_{ib}$ are generated both by person i and another person j, some external reciprocal relationships between i and j are the background for condition (2.7) or (2.6). However, we will focus first on person i's internal thoughts such as inferences/guesses from his own experiences, so we postpone our discussions about the background external relationships until Section 8.

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

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

$$\begin{aligned} D_{1a}^N &= \{(s_a, s_b^o) : s_a \in S_a\} \text{ and } D_{1b}^N = \emptyset \\ D_{2a}^N &= \emptyset \text{ and } D_{2b}^N = \{(s_a^o, s_b) : s_b \in S_b\}. \end{aligned} \tag{2.8}$$

With these domains, neither (2.6) nor (2.7) holds. Each person makes deviations over all his actions. However, each accumulates only active experiences, which means that he is either insensitive to (or ignores) the deviations by the other person. In this example, it would be natural, due to no role-switching, to assume that the frequency weights given as $\rho_{1a} = \rho_{2b} = 1$.

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

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

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

$$D_{1a}^A = \{(s_a, s_b^o) : s_a \in S_a\} \text{ and } D_{1b}^A = \{(s_a^o, s_b) : s_b \in S_b\}. \tag{2.9}$$

This means that person 1 makes trials with all actions for each role $r = a, b$, but he is insensitive to person 2's trials. If person 2 behaves in the same manner, then $D_{2a}^A = D_{1a}^A$ and $D_{2b}^A = D_{1b}^A$. Although both persons' domains are the same, the internal reciprocity condition (2.7) does not hold.

We give one domain that is internally reciprocal.

(3)(Reciprocal Active-Passive Domain): The *reciprocal active-passive domain* D_1^{AP} $= (D_{1a}^{AP}, D_{1b}^{AP})$ is given as:

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

Person 1 makes trials with all actions across both roles, and he is sensitive to both active and passive "unilateral" trials, but not joint-trials.[5] If person 2 has the same

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

personality, then 2 has the same domains: $D_{2a}^{AP} = D_{1a}^{AP}$ and $D_{2b}^{AP} = D_{1b}^{AP}$. This domain satisfies (2.7) and even (2.6), and is still smaller than the full reciprocal domain defined by $D_{ir}^{F} = S_a \times S_b$ for $i = 1, 2$, and $r = a, b$.

2.3. An Informal Theory of Behavior and Accumulation of Memories

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

We will refer to the above postulates in relevant places in this paper.

3. Direct and Transpersonal Understandings from Experiences

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

3.1. Transpersonal Postulates for the Other's Thoughts

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

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

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

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

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

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

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

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

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

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

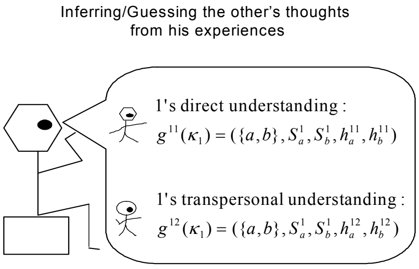
Inferring/Guessing the other's thoughts from his experiences
1's direct understanding :
$g^{11}(\kappa_1) = (\{a,b\}, S_a^1, S_b^1, h_a^{11}, h_b^{11})$
1's transpersonal understanding :
$g^{12}(\kappa_1) = (\{a,b\}, S_a^1, S_b^1, h_a^{12}, h_b^{12})$

Figure 3.1: Direct and Transpersonal Understandings

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

In the above metaphor, having a broken heart is an experience of losing a love, and it causes agony. Postulate TP1 requires that the agony caused by losing a love is understood by projecting one's past experience, which is $*1$. Then, postulate TP2 requires some experiential evidence (reason) to believe that she has broken heart. This is expressed as its contrapositive in $*2$: Since he has no experiential reason to believe her broken heart, he doubts her agony. This "reason to believe" is reminiscent of a requirement for the concept of "common knowledge" in Lewis [19]. In the next section, we will give an explicit formulation of the other's understanding based on postulates TP1 and TP2.

3.2. Direct and Transpersonal Understandings

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

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

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

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

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

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

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

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

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

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

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

These understandings are deterministic: All the components of $g^{ii}(\kappa_i)$ and $g^{ij}(\kappa_i)$, except θ_r for the unexperienced part of $S^i_a \times S^i_b$, are determined from the components of κ_i. This differs from in Kaneko-Kline [14], [15], and [16]. This determinism comes from our restriction on the 2-role game with assumptions Ob1 and Ob2.

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

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

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

The above requirement of having reason to believe is close to Lewis's [19] idea of person i having reason to believe that person j has also reason to believe the same. If we formulate the above argument as an epistemic logic system (cf., Kaneko [12]), we would examine this similarity more, which will be discussed in a separate paper. The argument here is entirely experiential, and in this sense, it is regarded as following the tradition from Mead [20].

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

(1)(Nonreciprocal Active Domain): Let (D_{1a}^N, D_{1b}^N) be given as the non-reciprocal domain of (2.8). In this example, person 1's d-understanding $g^{11}(\kappa_1)$ is given as: $S_a^1 = \{\mathbf{s}_{a1}, \mathbf{s}_{a2}, \mathbf{s}_{a3}\}$ and $S_b^1 = \{\mathbf{s}_{b1}\}$ by ID1[1]. Since person 1 has experiences for role a, the payoffs $(h_a^{11}(s_a, s_b), h_b^{11}(s_a, s_b))$ become those described in Table 3.1. Since person 1 has no experiences with role b, his understanding of those payoffs $h_b^{11}(s_a, s_b)$ is simply θ_b.

Table 3.1

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

Table 3.2

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

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

(2)(Reciprocal Active Domain): Let (D_{1a}^A, D_{1b}^A) be given by the active domain of (2.9). By ID1[1], we have $S_a^1 = \{s_{a1}, s_{a2}, s_{a3}\}$ and $S_b^1 = \{s_{b1}, s_{b2}, s_{b3}\}$. Then, it follows from ID2[11] that (h_a^{11}, h_b^{11}) is given as Table 3.3. When person 1 is at b, he cannot guess/infer his own payoffs from trials of person 2 at a. Thus, he puts θ_b to the payoffs from trials in the first column of Table 3.3. For the same reason, he puts θ_a in Table 3.3 along the top column. The remaining four strategy combinations (s_a, s_b) belong neither

D_{1a}^A nor D_{1b}^A, so he puts (θ_a, θ_b) in each case.

Table 3.3; g^{11}

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

Table 3.4; g^{12}

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

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

The above example may appear strange, which is due to the assumption that (D_{1a}^A, D_{1b}^A) is an active domain. When person 1 is equally sensitive to the experiences caused by person 2, he has the active-passive domain.

(3):(Reciprocal Active-Passive Domain): Let $D_1^{AP} = (D_{1a}^{AP}, D_{1b}^{AP})$ be the domains describe by (2.10). By ID^1, we have $S_a^1 = \{\mathbf{s}_{a1}, \mathbf{s}_{a2}, \mathbf{s}_{a3}\}$ and $S_b^1 = \{\mathbf{s}_{b1}, \mathbf{s}_{b2}, \mathbf{s}_{b3}\}$. But the payoff functions (h_a^{11}, h_b^{11}) are different from those in Table 3.3:

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

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

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

This internal reciprocity and coincidence will be important in our later analysis. We will give one theorem on this, which states that internal reciprocity (2.7) is necessary and sufficient for coincidence of a person's direct and transpersonal understandings up to the active and passive experiences.

Theorem 3.1 (Internal Coincidence). Tp-understanding $g^{ij}(\kappa_i)$ coincides with d-understanding $g^{ii}(\kappa_i)$ up to the active/passive experiences, i.e., $h_r^{ij}(s_a, s_b) = h_r^{ii}(s_a, s_b)$ for all $(s_a, s_b) \in \text{Proj}(S_a^i \times S_b^i)$ and all θ_a, θ_b if and only if (D_{ia}, D_{ib}) is internally reciprocal

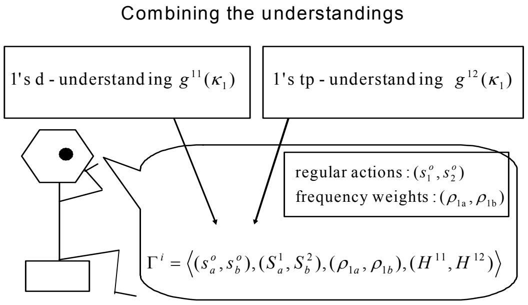
Combining the understandings
1's d - understanding $g^{11}(\kappa_1)$
1's tp - understanding $g^{12}(\kappa_1)$
regular actions : (s_1^o, s_2^o)
frequency weights : (ρ_{1a}, ρ_{1b})
$\Gamma^i = \langle (s_a^o, s_b^o), (S_a^1, S_b^2), (\rho_{1a}, \rho_{1b}), (H^{11}, H^{12}) \rangle$

Figure 4.1: The I.D.View

in the sense of (2.7).

Proof. **(If)**: Suppose $\text{Proj}(D_{ia}) = \text{Proj}(D_{ib})$. Hence, by (3.1) and (3.2), we have $h_r^{ij}(s_a, s_b) = h_r^{ii}(s_a, s_b)$ for all $(s_a, s_b) \in \text{Proj}(S_a^i \times S_b^i)$.

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

4. Inductively Derived Views and their Use for Behavioral Revision

4.1. Inductively Derived View

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

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

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

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

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

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

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 [25]. See Hu [10] for a more direct approach to expected utility theory from the frequentist perspective.

The definition of the i.d.view Γ^i has various differences from those given in Kaneko-Kline [14], [15] and [16]. One apparent difference is that the definition is given to a strategic game but not an extensive game (or an information protocol). This also makes the view here deterministic as $g^{ii}(\kappa_i)$ and $g^{ij}(\kappa_i)$. But it is the most important point to include the weighted payoffs coming from role-switching.

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

Now, consider how person i uses the i.d.view $\Gamma^i = \langle (s^o_a, s^o_b), (S^i_a, S^i_b), (\rho_{ia}, \rho_{ib}), H^{ii}, H^{ij} \rangle$. It includes the tp-understanding of the other person's payoffs in addition to his own d-understanding. When i uses Γ^i for his decision making, he would face the problem of whether or not he should use the tp-understanding. We have the following two cases:

C0(Partial Use): Person i uses only the payoff function H^{ii}, assuming that the other person j plays the regular action s^o_a or s^o_b in accordance with his assigned role.

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

In case C0, person i can maximize his weighted payoff H^{ii} by choosing his action from the assigned role in one play of the game. Since he uses only H^{ii}, he needs some assumption about the other person's action or reaction to his change. A simple assumption written in C0 is that the other person sticks to the regular action. In this case, person i may choose a maximum point against the regular action s^o_a or s^o_b. If both persons behave in this manner, or if the present regular actions are free from such behavior revisions, then the regular action pair (s_a^o, s_b^o) must be a Nash equilibrium in his understanding g^{ii}.

To see this, we consider only some reciprocal case with $\rho_{ia}, \rho_{ib} > 0$. Suppose that (s_a^o, s_b^o) is the regular behavior, and that person i considers a deviation at role a, say $[s_a, s_b^o]_a$. This is beneficial for him only if the weighted payoff

$$H^{ii}([s_a, s_b^o]_a, [s_a^o, s_b^o]_b) = \rho_{ia}h_a^{ii}(s_a, s_b^o) + \rho_{ib}h_b^{ii}(s_a^o, s_b^o) > \rho_{ia}h_a^{ii}(s_a^o, s_b^o) + \rho_{ib}h_b^{ii}(s_a^o, s_b^o).$$

This holds if and only if $h_a^{ii}(s_a, s_b^o) > h_a^{ii}(s_a^o, s_b^o)$. The same argument holds for role b. Hence, (s_a^o, s_b^o) is seen the be stable against such unilateral deviations if and only if it is a Nash equilibrium in person i's d-understanding. Since θ_r might appear in person i's d-understanding, a Nash equilibrium in g^{ii} may not be a Nash equilibrium in the base game G. The case C0 was discussed, without weighted payoff functions, in the context of an extensive game in Kaneko-Kline [14].

In the remainder of this paper, we study case C1. For this case, person i evaluates his action in terms of H^{ii}, and predicts what person j would do, by his H^{ij}. Suppose that

$$H^{ii}([s_a^o, s_b^o]_a, [s_a^o, s_b^o]_b) < H^{ii}([s_a, s_b^o]_a, [s_a, s_b^o]_b). \tag{4.3}$$

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

The answer is as follows: Suppose the parallel inequality:

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

If this holds, person i thinks that person j thinks in the same manner as (4.3) and shares in the benefit from this deviation. Person i now can believe that the deviation s_a from the regular action s_a^o gives a higher payoff for both persons, and thus that person j thinks in the same manner.

In (4.3) and (4.4), we considered only a unilateral derivation s_a from s_a^o, and we can also consider another parallel unilateral derivation s_b from s_b^o. Mathematically, we may consider even a joint deviation (s_a, s_b) from (s_a^o, s_b^o) satisfying (4.3) and (4.4). However, this requires some direct coordination or communication between the persons. In our context, we have a lot of possible ways of coordination or communication. It would be better to separate studies of these possibilities from the present research, which should be discussed in a separate paper.

Let us return to the unilateral deviation in (4.3) and (4.4). This deviation needs only one person to deviate first. That is, the deviation process can be expressed as

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

Fig.4.2

That is, suppose (s_a^o, s_b^o) is the current regular pair. Next, suppose that 1 deviates from s_a^o to a mutually beneficial s_a, which is the second left state in Fig.4.2. Then person 2 will observe this deviation, and when 2 is assigned role a, he follows 1's mutually beneficial deviation to s_a, which is describe as the third state in Fig.4.2.

5. Intrapersonal Coordination Equilibrium

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

In Section 4.2, we described the process of improving the weighted payoff for person i using his i.d.view Γ^i. In this section, first, we define a resulting equilibrium of this process. Let $\Gamma^i = \langle (s_a^o, s_b^o), (S_a^i, S_b^i), (\rho_{ia}, \rho_{ib}), H^{ii}, H^{ij} \rangle$ be the i.d.view derived from the memory kit κ_i.

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

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

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

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

That is, person i thinks, based on his i.d.view Γ^i, that s_a^o gives higher payoff to both persons 1 and 2 than any other action $s_a \in S_a^i$, and s_b^o has the same property.

The argument illustrated in Fig.4.2 is a specific case leading to the above definition. Inequalities (5.1) and/or (5.2) may include the case where we have the strict inequality for H^{ii} but the equality for H^{ij}; this may hold if person i's tp-understanding is trivial or poor. Thus, although the definition of an i.c.equilibrium is given by two inequality systems for each role, it includes cases where he has a poor tp-understanding or even his d-understanding is poor, e.g., S_a^i and S_b^i are small sets.

Our main target is an i.c.equilibrium for reciprocal cases, where we will have cooperation results. Nevertheless, we look at this case among other different cases in order to be able to discuss the conditions for cooperation to result. We start with cases of non-cooperative outcomes in Section 5, and discuss the cooperation results in Section 6.

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

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

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

5.2. Non-reciprocal Active Domain and Reciprocal Active Domain

There is a spectrum of reciprocal degrees of switching roles between the two persons. The non-reciprocal domains (2.8) and active domains (2.9) are located at the lowest side of this spectrum, while the fully reciprocal domains are located at the other extreme. It is our intention to show that cooperation is emerging as the reciprocal degree is increasing. To show this, we first show that at the lowest end, no cooperation occurs, more concretely, for the non-reciprocal domains and active domains, the i.c.equilibrium yields non-cooperative outcomes. In Section 6, we will consider the other extreme case of the spectrum of reciprocal degrees.

The first theorem is about the non-reciprocal domains.

Theorem 5.1 (Nonreciprocal Active Domain): Consider the non-reciprocal active domain (D_{ia}^N, D_{ib}^N) defined by (2.8) where person i takes role a. Then, the pair (s_a^o, s_b^o) of regular actions is an i.c.equilibrium in Γ^i if and only if it is a Nash equilibrium in person i's d-understanding g^{ii}.

Proof. For domains (D_{ia}^N, D_{ib}^N), person i's d-understanding g^{ii} is given as $S_r^i = S_r, S_{-r}^i = \{s_{-r}^o\}$ and $h_r^{ii}(s_r; s_{-r}^o) = h_r(s_r; s_{-r}^o)$, $h_{-r}^{ii}(s_r; s_{-r}^o) = \theta_{-r}$ for $s_r \in S_r$.

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

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

This implies $h_r(s_r^o; s_{-r}^o) \geq h_r(s_r; s_{-r}^o)$. Since role $-r$ has the unique choice, i.e., $S_{-r}^i = \{s_{-r}^o\}$, (s_a^o, s_b^o) is a Nash equilibrium in g^{ii}. Tracing the argument back, we have the only-if part, i.e., if (s_a^o, s_b^o) is a Nash equilibrium in g^{ii}, then (s_a^o, s_b^o) be an i.c.equilibrium in Γ^i. $\blacksquare$

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

Corollary 5.2 (Mutual I.C.Equilibrium in the Non-reciprocal Active Domains): Let D_i^{NA} be the non-reciprocal active domain with the regular actions (s_a^o, s_b^o) for $i = 1, 2$. Then, (s_a^o, s_b^o) is a mutual i.c.equilibrium if and only if it is a Nash equilibrium in the base game G.

We could obtain the corresponding Nash equilibrium results also for the non-reciprocal active-passive domain given in Section 2.2.

Next, we consider the reciprocal active domain and show that the insensitivity of persons having only active domains leads to a similar result even with a reciprocal role-switching.

Theorem 5.3 (Reciprocal Active Domain): Let (D_{ia}^A, D_{ib}^A) be the reciprocal active domain for person $i = 1, 2$, with the regular actions (s_a^o, s_b^o). Suppose that $\theta_a \leq h_a(s_a^o, s_b^o)$ and $\theta_b \leq h_b(s_a^o, s_b^o)$. Then the following two statements hold:

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

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

Proof. With this domain, the d-understanding $g^{ii} = (a, b, S_a^i, S_b^i, h_a^{ii}, h_b^{ii})$ is given by: for $r = a, b$, $S_r^i = S_r$, $h_r^{ii}(s_r; s_{-r}) = h_r(s_r; s_{-r})$ if $s_{-r} = s_{-r}^o$, and $h_r^{ii}(s_r; s_{-r}) = \theta_r$ otherwise. The tp-understanding $g^{ij} = (a, b, S_a^i, S_b^i, h_a^{ij}, h_b^{ij})$ is given as: for $r = a, b$, $S_r^i = S_r$, $h_r^{ij}(s_r; s_{-r}) = h_r(s_r; s_{-r})$ if $(s_r; s_{-r}) = (s_r^o; s_{-r}^o)$, and $h_r^{ij}(s_r; s_{-r}) = \theta_r$ otherwise.

(1): Let (s_a^o, s_b^o) be a Nash equilibrium in $G = (a, b, S_a, S_b, h_a, h_b)$. Consider role a and let s_a be an arbitrary action in S_a. Then, we have $h_a(s_a^o, s_b^o) \geq h_a(s_a, s_b^o)$. Thus, using the assumption that $\theta_b \leq h_b(s_a^o, s_b^o)$, we have

$$\begin{aligned} H^{ii}([s_a^o, s_b^o]_a, [s_a^o, s_b^o]_b) &= \rho_{ia} h_a^{ii}(s_a^o, s_b^o) + (1 - \rho_{ia}) h_b^{ii}(s_a^o, s_b^o) \\ &\geq \rho_{ia} h_a(s_a, s_b^o) + (1 - \rho_{ia})\theta_b = H^{ii}([s_a, s_b^o]_a, [s_a, s_b^o]_b). \end{aligned}$$

Now, the inequality for H^{ij} is:

$$\begin{aligned} H^{ij}([s_a^o, s_b^o]_a, [s_a^o, s_b^o]_b) &= \rho_{ia} h_b^{ij}(s_a^o, s_b^o) + (1 - \rho_{ia}) h_a^{ij}(s_a^o, s_b^o) \\ &\geq \rho_{ia}\theta_b + (1 - \rho_{ia})\theta_a = H^{ij}([s_a, s_b^o]_a, [s_a, s_b^o]_b). \end{aligned}$$

Thus, we have (5.1) of an i.c.equilibrium. Condition (5.2) can be shown in the same way letting s_b be an arbitrary action in S_b.

(2): Let (s_a^o, s_b^o) be an i.c.equilibrium in Γ^i. Then, we have

$$\begin{array}{rcl}\rho_{ia}h_a(s_a^o, s_b^o) + (1-\rho_{ia})h_b(s_a^o, s_b^o) & = & H^{ii}([s_a^o, s_b^o]_a, [s_a^o, s_b^o]_b) \\ & \geq & H^{ii}([s_a^o, s_b^o]_a, [s_a^o, s_b^o]_b) = \rho_{ia}h_a(s_a, s_b^o) + (1-\rho_{ia})\theta_b.\end{array}$$

Since $h_b(s_a^o, s_b^o) = \theta_b$, we have $h_a(s_a^o, s_b^o) \geq h_a(s_a, s_b^o)$. In the case of role b, we have a parallel argument, so (s_a^o, s_b^o) is a Nash equilibrium in Γ^i. ■

For the reciprocal active domains, we can have a corollary about a mutual i.c.equilibrium by combining the statements of Theorem 5.3 as in a similar manner to obtain Corollary 5.2 from Theorem 5.1. Even we may assume that one person has the non-reciprocal active domain and the other has a reciprocal active domain, and have a similar corollary. But this combination appears to be strange externally. External requirements will be given in Section 8. Here, it is enough to say that at the low end of the spectrum of reciprocal degrees, we have non-cooperative outcomes.

6. Intrapersonal Coordination Equilibrium for Reciprocal Domains

The results in Section 5.2 are noncooperative outcomes, since effectively due to lack of role-switching, or insensitivity to the other's trials, Γ^i does not describe anything about j's thinking. In contrast, when domains D_{ia} and D_{ib} are both reciprocal and the persons are sensitive to the other's trials, we would have very different results, which we will discuss in this section.

First, we will show the following result, which gives necessary conditions for an i.c.equilibrium. We emphasize that the frequency weights disappear in these conditions. These are reminiscent of *utilitarianism*, which has a different interpretation from "utilitarianism" in moral philosophy. This will be discussed in Section 7.

Theorem 6.1 (Utilitarian Criterion): Let (s_a^o, s_b^o) be an i.c.equilibrium for Γ^i with $(s_a^o, s_b^o) \in D_{ia} \cap D_{ib}$.

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

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

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

$$\begin{array}{rcl}\rho_{ia}h_a^{ii}(s_a^o, s_b^o) + (1-\rho_{ia})h_b^{ii}(s_a^o, s_b^o) & \geq & \rho_{ia}h_a^{ii}(s_a, s_b^o) + (1-\rho_{ia})h_b^{ii}(s_a, s_b^o); \quad (6.1) \\ \rho_{ia}h_b^{ij}(s_a^o, s_b^o) + (1-\rho_{ia})h_a^{ij}(s_a^o, s_b^o) & \geq & \rho_{ia}h_b^{ij}(s_a, s_b^o) + (1-\rho_{ia})h_a^{ij}(s_a, s_b^o).\end{array}$$

Since (s_a^o, s_b^o) and (s_a, s_b^o) are in $D_{ia} \cap D_{ib}$, it holds that for $r = a, b$,

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

Hence, summing up the first and second inequalities in (6.1), we have

$$h_a(s_a^o, s_b^o) + h_b(s_a^o, s_b^o) \geq h_a(s_a, s_b^o) + h_b(s_a, s_b^o).$$

■

When the persons are sensitive to the other's trials, and roles are switched as in the reciprocal active-passive domains, i.e., $D_{ia}^{AP} = D_{ib}^{AP} = \bigcup_{r=a,b}\{(s_r; s_{-r}^o) : s_r \in S_r\}$, parts (1) and (2) of Theorem 6.1 are simply the true payoff-sum maximization along the actions of each role. In the game of Table 2.1 with these domains, this payoff-sum maximization gives two possible candidates for an i.c.equilibrium: $(\mathbf{s}_{a2}, \mathbf{s}_{b1})$ and $(\mathbf{s}_{a1}, \mathbf{s}_{b2})$. Each of them together with $(D_{ia}^{AP}, D_{ib}^{AP})$ becomes an i.c.equilibrium. However, in the game of Table 2.2 with the affine transformation of role b's payoff, only $(\mathbf{s}_{a2}, \mathbf{s}_{b1})$ is an i.c.equilibrium. Thus we observe that an i.c.equilibrium is not invariant to such transformations of payoffs.

Theorem 6.1 gives necessary conditions for the resulting outcome of an i.c.equilibrium. The existence of an i.c.equilibrium in a reciprocal case is related to the degree of reciprocity in the frequency weights ρ_{ia}, ρ_{ib}. Although D_{ia}, D_{ib} are mathematically still independent of ρ_{ia}, ρ_{ib}, they should be closely related in interpretation: When D_{ia} and D_{ib} are reciprocal (not very different), so is (ρ_{ia}, ρ_{ib}) (not very different, too), and vice versa.

Now, we have the following theorem. In this theorem, a 2-role game $G = (a, b, S_a, S_b, h_a, h_b)$ is arbitrarily fixed. Only (D_{ia}, D_{ib}) and (ρ_{ia}, ρ_{ib}) must be specified. Recall that the internal reciprocity is defined by (2.7).

Theorem 6.2 (Existence of an I.C.Equilibrium 1): Let $(\rho_{ia}, \rho_{ib}) = (\frac{1}{2}, \frac{1}{2})$. Then, there is a pair $(s_a^o, s_b^o) \in S_1 \times S_2$ such that for any internally reciprocal domain (D_{ia}, D_{ib}) with $(s_a^o, s_b^o) \in D_{ia}$, the pair (s_a^o, s_b^o) is an i.c.equilibrium in Γ^i.

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

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

Since $S_a \times S_b$ is a finite set, we can find a pair $(s_a^o, s_b^o) \in S_a \times S_b$ satisfying (6.2).

Since (D_{ia}, D_{ib}) is internally reciprocal, it follows from Theorem 3.1 and (2.3) that for all $(s_a, s_b) \in S_a^i \times S_b^i$,

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

Hence, using (6.2), we have, for all $s_a \in S_a^i$ and $s_b \in S_b^i$,

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

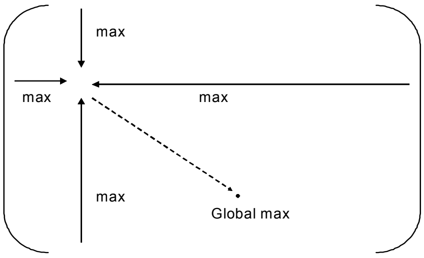

Figure 6.1: Candidates for an i.c.equilibria

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

In the above proof, the pair (s_a^o, s_b^o) chosen by (6.2) reaching the maximum payoff sum is independent of person i. Hence, Theorem 6.1 can be read for both persons with a common point (s_a^o, s_b^o). Hence, (s_a^o, s_b^o) is a mutual i.c.equilibrium. We state this fact as a corollary.

Corollary 6.3 (Existence of a Mutual I.C.Equilibrium): Let $(\rho_{ia}, \rho_{ib}) = (\frac{1}{2}, \frac{1}{2})$ for $i = 1, 2$. Then, there is a pair $(s_a^o, s_b^o) \in S_1 \times S_2$ such that for any internally reciprocal domain (D_{ia}, D_{ib}) with $(s_a^o, s_b^o) \in D_{ia}$ for $i = 1, 2$, the pair (s_a^o, s_b^o) is a mutual i.c.equilibrium.

In the proof of Theorems 6.2, the pair (s_a^o, s_b^o) is chosen as a global maximization point over the entire matrix. But a necessary choice is made over the set such as the one described in Fig.6.1. Once this is recognized, a simple algorithm to such a point is found: Take any pair in the matrix. Then, if there is one pair with a higher weighted sum of payoffs obtained by one person's deviation, we move to this pair. If this pair has the same property, then we move again. Then, we will reach one pair without a further improvement. This convergence holds since the matrix is finite and each step has an improvement in the weighted sum of payoffs. The resulting pair may not be a global maximization point.

Theorems 5.1 and 5.2 suggest that an i.c.equilibrium may not exist in the nonreciprocal cases with insufficient role-switching. The following example shows possible nonexistence, even under $D_{1a} = D_{1b}$, without $\rho_{ia} = \rho_{ib} = 1/2$.

Example 6.1: Consider the game of Table 2.1. Suppose that D_{1a} and D_{1b} are the active-passive domain given by (2.10) with $(s_a^o, s_b^o) = (\mathbf{s}_{a2}, \mathbf{s}_{b1})$, and $\rho_{ia} = \rho_{ib} = 1/2$. Then, $(\mathbf{s}_{a2}, \mathbf{s}_{b1})$ is an i.c.equilibrium. In this case, we can choose the full set $S_1 \times S_2$

or the smallest set $\{(\mathbf{s}_{a2}, \mathbf{s}_{b1})\}$ as D_{1a} and D_{1b} without breaking this i.c.equilibrium. In the same way the alternative regular behavior $(s_a^o, s_b^o) = (\mathbf{s}_{a1}, \mathbf{s}_{b2})$ is an i.c.equilibrium across the same active-passive domain.

For the payoffs of Table 2.2, however, only $(\mathbf{s}_{a2}, \mathbf{s}_{b1})$ is an i.c.equilibrium. Thus, this equilibrium concept is not independent of a positive linear transformation of a the payoff function of a specific role.

Let us return to the game of Table 2.1 with the active-passive domain D_{1a} and D_{1b} with $(s_a^o, s_b^o) = (\mathbf{s}_{a2}, \mathbf{s}_{b1})$ and also containing $(\mathbf{s}_{a1}, \mathbf{s}_{b1})$. When the frequency weights are different, it may not have an i.c.equilibrium. Let $\rho_{ia} = 9/10$ and $\rho_{ib} = 1/10$. Then, we have

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

Hence, $(\mathbf{s}_{a2}, \mathbf{s}_{b1})$ is not an i.c.equilibrium. The other candidate for an i.c.equilibrium is $(\mathbf{s}_{a1}, \mathbf{s}_{b2})$, but this is not an equilibrium either. Thus, with this weight, the assertion of Theorem 6.2 does not hold.

When $\rho_{ia} = 1/3$ and $\rho_{ib} = 2/3$, $(\mathbf{s}_{a2}, \mathbf{s}_{b1})$ becomes again an i.c.equilibrium, though the necessary conditions given by Theorem 6.1 remains. Thus, we have some interval containing ρ_{ia} so that for weights in the interval, we have an i.c.equilibrium.

The above example indicates that even though an i.c.equilibrium may disappear for some (ρ_{ia}, ρ_{ib}), it may remain for (ρ_{ia}, ρ_{ib}) close to $(\frac{1}{2}, \frac{1}{2})$. This fact is relevant, as stated in Section 2.1, for our interpretation that the values of weights ρ_{ia} and ρ_{ib} should not interpreted as exact values. The following variant of Theorem 6.2 shows that this interpretation has some legitimacy. We state the theorem without a proof.

Theorem 6.4 (Existence of an I.C.Equilibrium 2): Let (s_a^o, s_b^o) be a pair so that

$$\begin{aligned}h_a(s_a^o, s_b^o) + h_b(s_a^o, s_b^o) &> h_a(s_a, s_b^o) + h_b(s_a, s_b^o) \text{ for all } s_a \in S_a^i \backslash \{s_a^o\} \\ h_a(s_a^o, s_b^o) + h_b(s_a^o, s_b^o) &> h_a(s_a^o, s_b) + h_b(s_a^o, s_b) \text{ for all } s_b \in S_b^i \backslash \{s_b^o\}.\end{aligned}$$

Then there is a $\alpha \in (0, \frac{1}{2})$ such that for any $\rho_{ia} \in (\frac{1}{2} - \alpha, \frac{1}{2} + \alpha)$ and any internally reciprocal (D_{ia}, D_{ib}) with $D_{ia} \ni (s_a^o, s_b^o)$, (s_a^o, s_b^o) is an i.c.equilibrium in Γ^i.

We could also consider conditions for existence taking into account θ_a and θ_b as well as ρ_{ia}, ρ_{ib}. Since, however, this leads us further astray from our main intentions, we do not pursue this here.

7. Applications to the Ultimatum Game and Dictator Game

Here, we apply the results of Section 6 to the ultimatum game and dictator game. For those games, experimental results differ consistently from the non-cooperative theoreti-

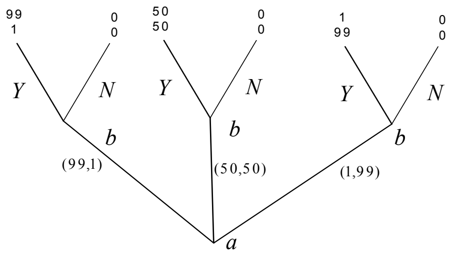

Figure 7.1: Ultimatum Game

cal predictions based on the standard equilibrium theory. Cooperative outcomes (equal division) result more often in experiments than the predicted non-cooperative outcomes (cf. Güth *et al.* [7], Kahneman *et al.* [11] and also Camerer [3]). Here, we consider simplified variants of those games, and apply our theory to them.

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

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

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

Under the assumption that each $i = 1,2$ has the reciprocal active-passive domain $D^{AP}_{ia} = D^{AP}_{ib}$ and $\rho_{1a} = \rho_{2a} = 1/2$, a pair $((99,1),(Y,Y,Y))$ is not an i.c.equilibrium since

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

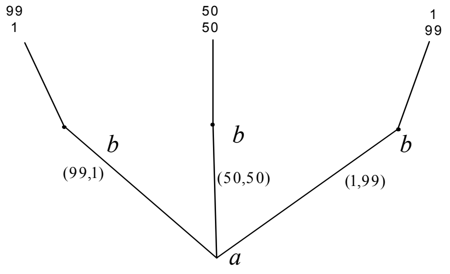

Figure 7.2: Dictator Game

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

There are several issues here. One is that we have treated this game as a strategic game to fit into the theory given in this paper. In order to study it as an extensive game, we need to extend our theory to extensive games or information protocols such as in [14] and [15]. Another issue is that the i.c.equilibrium does not consider joint deviations, so equilibria like $((99,1),(Y,N,N))$ can persist. As mentioned in Section 4.2, we could have extended our theory to include joint deviations, but chose not to do so, since it would include other conceptual problems. With such an extension, we could discuss how the other equilibria such as $((99,1),(Y,N,N))$ may or may not remain in our theory.

Here, instead of extending the present theory, we simplify the ultimatum game so as to treat it as a strategic game to show how the results become more clear cut. We will treat a simpler version of the dictator game given by Kahneman *et al.* [11] (see also Camerer [3] for a survey of experimental studies of dictator games).

Dictator Game: Let us eliminate action N from each move of role b. The game is depicted as Fig.7.2. This has no action choice for role b, and thus, it is regarded a 1-role game from the standard game theoretic point of view. However, payoffs to role b matter in our theory. First, we consider:

Case 1: Reciprocal Active-Passive Domains: Here, we first specify the domain and frequencies of role-switching with $(s_a^o, s_b^o) = ((50,50),Y)$: for $i = 1,2$,

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

Then $S_a^i = \{(99,1),(50,50),(1,99)\}$ and $S_b^i = \{Y\}$ for $i = 1,2$. Here, we have the unique i.c.equilibrium $((50,50),Y)$.

If we apply the Nash equilibrium concept, then the person assigned to role a chooses $(99,1)$ to maximize his payoff (utility) and to receive 99, and the one at role b should simply follow this and receive payoff 1. This is not an i.c.equilibrium: Indeed, when they switch the roles, one person obtains \$99 and \$1 with frequencies 1/2 and 1/2, respectively. This alternating payoffs are less preferred to taking \$50 constantly, since the utility function u is strictly concave.

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

Case 2: Non-reciprocal Active Domains:

$$\begin{aligned} D_{1a} &= \{((99,1),Y),((50,50),Y),((1,99),Y)\} \text{ and } D_{1b} = \emptyset \\ \rho_{1a} &= 1 \text{ and } \rho_{2b} = 1. \end{aligned} \tag{7.2}$$

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

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

Discussions of the Above Results: Social Contexts:

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

Since the dictator game and results for it are simpler than the results for the ultimatum game, we should focus on the former. One possible hypothesis is that the fully reciprocal Case 1 with equal sharing corresponds to the standard experimental design where the roles and the opponents are chosen randomly in each round keeping their anonymity. This experimental design already captures our internal reciprocity well and the experimental results of sharing fit well.

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

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

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

Now we turn to morality or fairness. It is our contention that as far as a situation is recurrent and reciprocal enough, the persons possibly cooperate in the form of the simple payoff sum maximization. Since this is, perhaps, quite pervasive for human relations among small numbers of people, they could have such patterned behavior, and consequently, such behavior is then observed in experiments. This gives an "anthropological", i.e., "experiential" grounding for morality. This differs from the rationalistic school of morality - - it comes from rationalistic reasoning about morality (such as in Harsanyi [8]). It also differs from Adam Smith's [24] "moral sentiments" - - people are born with a moral sense. In our case, the "morality" of the form of the payoff-sum maximization emerges from social interactions and role-switching in complex social web, and is neither rationalized nor inborn. We regard this as an anthropological foundation for the "utilitarianism" expressed in the form of Theorem 6.1.

8. Externally Reciprocal Relations

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

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

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

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

That is, if person j has a passive experience, then person i must have this as an active experience causing j's passive experience. This is of the same nature as the Postulates EP3 and EP4 of Section 2. Based on these postulates, (8.1) formulates the idea that a person is more sensitive to being active with respect to memories. This gives an element of reciprocity but is only a necessary form of reciprocity. For example, the non-reciprocal active domains D_1^N and D_2^N still satisfy (8.1).

Each person may be more sensitive to his own deviations; as time is going on, he may have learned also passive experiences. Eventually, the converse of (8.1) could hold:

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

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

That is, (8.1) becomes the equivalence between both sides.

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

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

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

Example. 8.1 (Different Trials): Consider the game in Table 2.1 and the following

D_1, D_2 with the regular actions $(s_a^o, s_b^o) = (\mathbf{s}_{a1}, \mathbf{s}_{b1})$;

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

That is, person 1 makes trials of only the second actions s_{a2} at role a and s_{b2} at role b, while person 2 makes trials of only the third actions s_{a3} at a and s_{b3} at b. Even though (8.2) holds, and their roles are switched, their differences in trial behaviors generates different domains of experiences, i.e., $D_{ia} \neq D_{ib}$ for $i = 1, 2$ and $D_{1r} \neq D_{2r}$ for $r = a, b$, though $D_{1a} = D_{2b}$ and $D_{2a} = D_{1b}$. These inequalities prevent them from constructing meaningful tp-understandings. This fact implies that regardless of the weights ρ_1, ρ_2, that if $(\mathbf{s}_{a1}, \mathbf{s}_{b1})$ is an i.c.equilibrium and $\theta_r = h_r(\mathbf{s}_{a1}, \mathbf{s}_{b1})$ for $r = a, b$, then it is a Nash equilibrium in the restricted game $(a, b, \{\mathbf{s}_{a1}, \mathbf{s}_{a2}\}, \{\mathbf{s}_{b1}, \mathbf{s}_{b3}\}, h_a, h_b)$.

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

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

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

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

Recalling $\text{Proj}(D_{ir}) := \{(s_a, s_b) \in D_{ir} : s_a = s_a^o \text{ or } s_b = s_b^o\}$, we can change (8.2) and (8.3) to equivalent but mathematically clearer conditions.

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

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

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

Proof. When (1) and (2) hold, the four sets, $\text{Proj}(D_{ir})$, $i = 1, 2$ and $r = a, b$ are the same. Hence, the if-part is straightforward. We prove the only-if part. Suppose (8.2) and (8.3) for (D_{1a}, D_{1b}) and (D_{2a}, D_{2b}).

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

Consider (2). Let $(s_a, s_b) \in \text{Proj}(D_{1a})$. This means that $(s_a, s_b) = (s_a, s_b^o)$ or (s_a^o, s_b). Let $(s_a, s_b) = (s_a, s_b^o)$. By (8.3), we have $(s_a, s_b^o) \in \text{Proj}(D_{2a})$. 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 $(s_b; s_{-b}^o) \in$

Proj(D_{1b}). By (8.2), we have $(s_b; s^o_{-b}) \in$ Proj(D_{2a}). We have shown that Proj$(D_{1a}) \subseteq$ Proj(D_{2a}). The converse can be obtained by a symmetric argument. Thus we have (2). ■

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

In Theorem 2.1, we have already seen that internal reciprocity (1) is necessary and sufficient for g^{ii} and g^{ij} to coincide within the mind of one person i. The next step is to consider when the two persons reach the same views. In this case, under the assumption of $\rho_{1a} = \rho_{2a} = 1/2$, a mutual i.c.equilibrium makes sense.

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

Theorem 8.2.(Internally and Externally Reciprocal Relations): Then, (8.2) and (8.3) hold for (D_{1a}, D_{1b}) and (D_{2a}, D_{2b}) if and only if for any $r = a, b$ and $i, j = 1, 2$ $(i \neq j)$,

(1): $S^i_r = S^j_r$;

(2): for any $(s_a, s_b) \in$ Proj$(S^1_a \times S^1_b)$ and θ_a, θ_b, $h^{ii}_r(s_a, s_b) = h^{ij}_r(s_a, s_b) = h_r(s_a, s_b)$.

Proof. **(Only-If)**: Suppose that (8.2) and (8.3) hold for (D_{1a}, D_{1b}) and (D_{2a}, D_{2b}). Then, Lemma 8.1 states that Proj(D_{ir})'s are all the same for $i = 1, 2$ and $r = a, b$. Hence, (1) is satisfied by the d-understanding $g^{ii} = (a, b, S^i_a, S^i_b, h^{ii}_a, h^{ii}_b)$ and tp-understanding $g^{ij} = (a, b, S^i_a, S^i_b, h^{ij}_a, h^{ij}_b)$ for $i, j = 1, 2$ $(i \neq j)$. Assertion (2) also follows by (3.2) and (3.1).

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

An implication of Theorem 8.2 is that under (8.2), (8.3) and the frequency assumption that $(\rho_{ia}, \rho_{ib}) = (\frac{1}{2}, \frac{1}{2})$ for $i = 1, 2$, person i can predict the correct payoff function over the relevant domains, that is, for any $s_r \in S^j_r = S^i_r$,

$$H^{ij}([s_r; s^o_{-r}]_a, [s_r; s^o_{-r}]_b) = H^{jj}([s_r; s^o_{-r}]_a, [s_r; s^o_{-r}]_b). \tag{8.4}$$

Hence, those persons think about the game in the perfectly synchronized manner, *a fortiori*, if (s^o_a, s^o_b) is a mutual i.c.equilibrium, then they reach the understanding that it is an i.c.equilibrium for both persons.

We do not claim that these two types of reciprocities are reached even after the situation has been played with full role-switching. Here, we give only one example where each has internally reciprocal domains but they are not externally reciprocal

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

In this case, the dynamics suggested in Fig.5.2 may not work. For example, person 1 thinks that a deviation s_a gives a better weighted payoff, and he thinks that person 2 thinks in the same manner. But, if the experience (s_a, s_b^o) is not accumulated in person 2's mind, person 2 does not deviate as 1 predicts. In this case, person 1 may find that person 2's i.d.view is different.

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

9. Conclusions

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

9.1. Summary of Findings

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

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

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

Nevertheless, with different degrees of reciprocity, we have many cases for the domains of accumulation generated, some of which were well suited for cooperation, and others not. The reciprocal active-passive domain was shown to be well suited for cooperation to emerge in an i.c.equilibrium (Theorem 6.2). On the other hand, the non-reciprocal and reciprocal active domains generate only non-cooperative Nash equilibria as i.c.equilibria (Theorems 5.2 and 5.3). In this paper, we pursued only a few cases to express the main thrust of the arguments about the potential for the emergence of cooperation.

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

Section 8 gave external conditions for the internal reciprocity which was at the heart of the emergence of cooperation in our theory. This exploration showed that in addition to reciprocal role-switching, the same trials by both persons, and equal and broad sensitivities were sufficient (Theorem 8.2) to generate the equivalent understandings that are fertile grounds for cooperation (Theorem 6.2).

9.2. Extensions and Future Work

First, we discuss some implicit assumptions underlying of formulation of person i's derivation of the tp-understanding about the other's understanding of the situation. It is experiential in the sense that all components are derived from his own accumulated experiences. Here, we need social roles, role-switching, and also the basic assumption that the 2-role game is given independent of persons (actors). In this sense, we have followed the tradition of symbolic interactionism from Mead [20]. In reality, this could not be true, but our assumption is an idealization. Not only this, we need other assumptions for our treatment. These were specified from place to place in this paper.

Specifically, the definition of person i's tp-understanding g^{ij} from his memory kit κ_i includes such an assumption. The salient part is the condition $(s_a, s_b) \in D_{i(-r)}$ in the definition of $h_r^{ij}(s_a, s_b)$ in (3.2). This means that person i has the experience (s_a, s_b) in his domain $D_{i(-r)}$: He infers that person j must have also this experience, and project his experienced payoff $h_r(s_a, s_b)$ onto j. That is, (s_a, s_b) must be a common experience for persons i and j from the viewpoint of person i. This sounds like the requirement of some evidence for the definition of common knowledge in Lewis [19]. This will possibly serve a bridge to epistemic logic.

In our context, if a person experiences one pair (s_a, s_b) from both roles, he would infer/guess that the other person has the same experience from both roles, and also that the other infer the symmetric statement. If we pursue, rigorously, this argument as an infinite regress, then we would have common knowledge (beliefs) in the sense of an infinite hierarchy of beliefs (see Kaneko [13], Chap.4 for this argument of an infinite regress). In this case, we need other assumptions on an hierarchy of logical abilities of the persons. As Lewis [19] did not intend to mean an infinite hierarchy of knowledge (beliefs), it would be better to stop at some shallow interpersonal depths of nested beliefs. We will discuss a rigorous treatment of this in the epistemic logic of shallow depths (Kaneko-Suzuki [18]) in a separate paper.

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

In an n-person case, since one person can experience a few social roles only such as in a baseball game, it would be not important to extend directly the result of this paper into the n-person case. Rather we should consider possibilities of cooperations of 2- or 3-person groups in the entire n-person game. Hence, our research does not suggest us to return to the standard cooperative game theory from von Neumann-Morgenstern [25].

Rather, patterned behavior in different but similar situations may be a key to have an extension of our theory. This is related to the basic presumption of inductive game theory: A social situation formulated as a 2-role game (more generally, an n-role game) is not isolated from other social situations in the entire social web. As a research strategy, we focus on a specific 2-role game in this paper, but it belongs to the social web depicted as Fig.1.1. Also, our behavioral postulate is a patterned behavior, rather than instantaneous payoff maximization. This patterned behavior have some uniformity (regularity), which may ease some difficulty to reach cooperation such as one problem difficulty we met in the ultimatum game in Section 7. This thought may suggest more experiential studies of behavior in society and experimental studies in labs.

References

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

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

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

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

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

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

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

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

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

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

[11] Kahneman, D., J. L. Knetsch and R. Thaler, (1086), Fairness as a Constraint on Profit Seeking: Entitlements in the Market, *American Economic Review* 76, 728-741.

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

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

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

[15] Kaneko, M., and J. J. Kline, (2008b), Information Protocols and Extensive Games in Inductive Game Theory, *International Journal of Mathematics, Game Theory and Algebra 17,* issue *5/6*, 2008.

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

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

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

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

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

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

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

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

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

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