Undecidability arising from Prediction/Decision Making in an Interdependent Situation[*]

Tai-Wei Hu[†]and Mamoru Kaneko[‡]

09 August 2014, preliminary

Abstract

Logical inference is an engine for human thinking, especially, for decision making in an interdependent situation with more than one persons. We study the possibility of prediction/decision making in a finite 2–person game with pure strategies, following the Nash-Johansen noncooperative solution theory. Since some infinite regress naturally arises in this theory, we adopt a fixed-point extension IR^2 of the epistemic logic KD^2, which is still a finitary propositional logic. The base logic KD^2 is adopted to capture individual decision making from the viewpoint of logical inference. Our results differ between a game with the interchangeable set of Nash equilibria and a game with the uninterchangeable set. For the former, we have decidability, i.e., player i can decide whether each of his strategies is a final decision or not. For the latter, he can neither decide it to be a possible decision nor can disprove it. This takes the form of Gödel's incompleteness theorem, while it is much simpler. Our undecidability also is related to the self-referential structure, but its main source is interdependence of payoffs and independent prediction/decision making.

Key words: Prediction/Decision Making, Infinite Regress, Decidability, Undecidablity, Incompleteness, Nash solution, Nash subsolution

1 Introduction

Logical inference is an engine for decision making in complex situations, in particular, in interdependent situations with multiple persons like games. Decision making in such situations has been studied in game theory, while logical inference is kept informal. To study this decision making, we adopt a formal system of an epistemic logic; specifically, a fixed-point extension IR^2 of the (propositional) epistemic logic KD^2. Since prediction making is also required because of interdependence of the players, it is more accurate to call "prediction/decision making". Nash [17] and Johansen [10] gave the noncooperative theory of prediction/decision making in a non-formalized manner. We study this theory in the logic IR^2.

We prove the undecidability (incompleteness) result that for a game with the uninterchangeable set of Nash equilibria, a player may reach neither a positive nor a negative decision; i.e.,

[*] The authors are partially supported by Grant-in-Aids for Scientific Research No.26234567 and No.2312002, Ministry of Education, Science and Culture.

[†] Northwestern University, Illinois, USA, t-hu@kellogg.northwestern.edu

[‡] Faculty of Political Science and Economics, Waseda University, Tokyo, Japan, kaneko@sk.tsukuba.ac.jp

the belief set for him is incomplete in the logic IR^2. In contrast, when the set of Nash equilibria is interchangeable, the same belief set leads him to decidability.

Our approach has various different features from the standard literatures of game theory as well as epistemic logic. Thus, we start with explaining those features.

Fixed-point extension of $\mathbf{KD}^2$: As seen presently, we meet an infinite regress of beliefs of prediction/decision making by one player and about the other player. This infinite regress occurs in the mind of a single player, i, and we need to separate one mind from the other. In order to keep subjective thinking separately from the other's, we adopt the epistemic logic KD^2 as the base logic for IR^2. This separation has several merits, which will be explained in several places.

The concept of an infinite regress is closely related to the common knowledge (Lewis [14] and Aumann [1]), and the logic IR^2 is related to the common knowledge logic CKL (cf., Fagin, *et al.* [5], and Meyer-van der Hoek [15]). Indeed, if we assume Axiom T (truthfulness) on IR^2, infinite regress collapse to common knowledge, and IR^2 becomes equivalent to CKR.

Proof theory and model theory: The view of regarding logical inference as an engine for prediction/decision making leads us to a proof-theoretic system. The status of model theory (Kripke semantics) is rather a technical support, though we need it to stabilize the choice of a formal system. Indeed, we prove the Kripke-soundness/completeness of IR^2 in a separate paper, [8]. We emphasize that a player's prediction/decision making is formulated in a proof theoretic (formal) system, and rather than in a single (semantic) model[1,2]. The Kripke completeness for IR^2 tells that semantic validity is captured by formal provability, and soundness tells that a counter model disproves provability. The soundness part will be used for our undecidability theorem. but the status of model theory is still secondary for our study.

In the logic IR^2, the logical ability of each player consists of the ability given by classical logic and the knowledge about the same ability of the other, and additionally, the ability manipulating infinite regresses.

Basic beliefs as non-logical axioms: As a mathematical theory needs its proper mathematical axioms in a formal system, a player's prediction/decision making needs *basic beliefs* (understanding) of the *situation* (game) and his *prediction/decision criterion*; otherwise, he could only recognize logical deducibility. The deducibility from his beliefs to a decision is expressed as

$$\mathbf{B}_i(\Gamma_i^o) \vdash \mathbf{B}_i(\mathrm{I}_i(s_i)). \tag{1}$$

That is, player i has basic beliefs Γ_i^o in his mind, and derives $\mathrm{I}_i(s_i)$ - "s_i is a possible decision for him". The negative decision is described by $\mathbf{B}_i(\Gamma_i^o) \vdash \mathbf{B}_i(\neg\mathrm{I}_i(s_i))$. In the logic IR^2, $\mathbf{B}_i(\Gamma_i^o) \vdash \mathbf{B}_i(\mathrm{I}_i(s_i))$ (respectively, $\mathbf{B}_i(\Gamma_i^o) \vdash \neg\mathbf{B}_i(\mathrm{I}_i(s_i))$ is equivalent to $\Gamma_i^o \vdash \mathrm{I}_i(s_i)$ ($\Gamma_i^o \vdash \neg\mathrm{I}_i(s_i)$). This is interpreted as meaning that the derivation of (1) is done in the mind of player i. For this, our choice of the base logic KD^2 is essential (see Lemma 2.5).

Game theoretical concepts: We consider only finite 2-person (strategic) games with pure strategies. This simple setting is enough for our considerations of undecidability for predic-

[1] The model-theoretic standpoint has been taken almost exclusively in the literature of epistemic logic with applications to game theory; for example, see van Benthem *et al.* [22], Perea [19], in the various papers in Brandenbuger [4], and van Benthem [21]. Some exceptions are Kaneko-Nagashima [11], Kline [13], and Suzuki [20], where the proof-theoretic standpoint is taken.

[2] Many aspects involved in playing a game are considered in van Benthem *et al.* [22] and van Benthem [21]. In Chap.12 of [21], matrix games are considered from the viewpoint of logic; matrix games are formulated by means of logic. Neverrtheless, an individual thought process of prediction/decision making is only indirectly treated.

tion/decision making.

Nash [17] distinguished between a *solvable* game and an *unsolvable* game, which will be explained in Section 3. His theory is well suited to an interpretation of individual *ex ante* prediction/decision making in a game, but he stopped at giving the distinction. Johansen [10] discussed Nash's theory in a more philosophical manner. He focussed on solvable games. Our axiomatic description of prediction/decision making may be regarded as a formalization of his argument in the formal system IR^2.

Axiomatic formulation of prediction/decision making: We formulate prediction/decision making by three axioms N0$_i$, N1$_i$, and N2$_i$, given in Section 4. They are assumed in the scope of the mind of player i, i.e., $\mathbf{B}_i(\text{N012}_i) := \mathbf{B}_i(\text{N0}_i\wedge\text{N1}_i\wedge \text{ N2}_i)$. For his prediction about the other player j's decision making, player i should have beliefs $\mathbf{B}_i\mathbf{B}_j(\text{N012}_j)$, where N012$_j$ is the same as N012$_i$ with the replacement of i with j. In fact, $\mathbf{B}_i\mathbf{B}_j(\text{N012}_j)$ requires $\mathbf{B}_i\mathbf{B}_j\mathbf{B}_i(\text{N012}_i)$, and so on. This regress generates an infinite sequence:

$$\mathbf{B}_i(\text{N012}_i),\ \mathbf{B}_i\mathbf{B}_j(\text{N012}_j),\ \mathbf{B}_i\mathbf{B}_j\mathbf{B}_i(\text{N012}_i), ... \qquad (2)$$

This infinite regress is captured by the fixed-point operator as $\mathbf{Ir}_i(\text{N012}_i;\text{N012}_j)$ in the logic IR^2. It has a self-referential structure; i.e., player i has the imaginary j, and this j has also the imaginary i in his mind, and *vice versa.* The self-referential structure is crucial for our undecidability result.

The infinite regress $\mathbf{Ir}_i(\text{N012}_i;\text{N012}_j)$ describes a property for prediction/decision making, but there are multiple candidates to enjoy this property. Among them, we choose a candidate formula to have exactly the property, which is formulated as $\mathbf{Ir}_i(\mathbf{WF})$.

Infinite-regress of preferences: The set of beliefs $\mathbf{Ir}_i(\text{N012}_i;\text{N012}_j), \mathbf{Ir}_i(\mathbf{WF})$ is a pure description of how player i makes predictions and decisions, but yet it includes no concrete information about a game. The corresponding beliefs of game payoffs are described as $\mathbf{Ir}_i(g_i; g_j)$. We adopt the set of those three types of beliefs, $\Delta_i = \{\mathbf{Ir}_i(g_i; g_j), \mathbf{Ir}_i(\text{N012}_i;\text{N012}_j)\} \cup\mathbf{Ir}_i(\mathbf{WF})$. We consider two related questions:

(i): What decisions and predictions does Δ_i recommend?

(ii): Does it, in the first place, recommend any?

These are related closely to Nash's [17] distinction between solvable and unsolvable games. Here, interchangeability is more directly related.

Table 1.1

	s$_{21}$	s$_{22}$	s$_{23}$
s$_{11}$	$2,4$	$2,2$	$4,0$
s$_{12}$	$3,3^{NE}$	$4,2$	$3,0$
s$_{13}$	$0,0$	$5,5$	$2,6$

Table 1.2

	s$_{21}$	s$_{22}$
s$_{11}$	$2,1^{NE}$	$0,0$
s$_{12}$	$0,0$	$1,2^{NE}$

Table 1.3

	s$_{21}$	s$_{22}$
s$_{11}$	$1,-1$	$-1,1$
s$_{12}$	$-1,1$	$1,-1$

Interchangeable and uninterchangeable games: In (the game of) Table 1.1, each player has three strategies, and his payoff is determined in the matrix (the first entry is player 1's payoff). The superscript NE stands for Nash equilibrium, explained in Section 3. Table 1.1 has a unique Nash equilibrium. Table 1.2 has two Nash equilibria. This has the uninterchangeable set of Nash equilibria in the sense that a pair of strategies from those equilibria may not be an equilibrium. Table 1.3 has the empty set of Nash equilibria. The sets of Nash equilibria for Tables 1.1 and 1.3 are interchangeable. Table 1.3 has a difficulty of no Nash equilibrium, which is different from the difficulty caused by undecidability. We say that a game is (*un-*) *interchangeable* if the set of Nash equilibria is (un-) interchangeable.

Positive decision, negative decisions, and undecidable: Both questions (i) and (ii) are related to Nash's distinction. When (the set of Nash equilibria for) a game is interchangeable such as in Tables 1.1 and 1.3, we have the following decidability result: for *any* strategy s_i for player i,

$$\text{either } \mathbf{\Delta}_i \vdash \mathbf{B}_i(\mathrm{I}_i(s_i)) \text{ or } \mathbf{\Delta}_i \vdash \mathbf{B}_i(\neg \mathrm{I}_i(s_i)). \tag{3}$$

Furthermore, decision $\mathrm{I}_i(s_i)$ can be expressed as a concrete formula. The set of beliefs $\mathbf{\Delta}_i$ tells that in Table 1.1, $\mathbf{s}_{12}$ is a positive decision but both $\mathbf{s}_{11}$ and $\mathbf{s}_{13}$ are negative decisions. In Table 1.3, $\mathbf{\Delta}_i$ recommends all as negative decisions.

Now, it is the main result of the paper that when a game is uninterchangeable such as Table 1.2, there is *some* strategy s_i for each player i such that

$$\text{neither } \mathbf{\Delta}_i \vdash \mathbf{B}_i(\mathrm{I}_i(s_i)) \text{ nor } \mathbf{\Delta}_i \vdash \mathbf{B}_i(\neg \mathrm{I}_i(s_i)). \tag{4}$$

That is, player i cannot decide with the same belief set $\mathbf{\Delta}_i$ whether s_i is a positive or negative decision. This holds for both strategies in Table 1.2. This entirely differs from the case where he finds all as negatively recommended, since if so, he may look for a different criterion. However, in the case of (4), he may not be able to notice this undecidability itself.

Relations to Gödel's incompleteness theorem and the source for our undecidability: The result (4) has the same form as Gödel's incompleteness theorem (cf., Boolos [3], Mendelson [16]), but both interpretation and source for incompleteness are different. Gödel's theorem is about the Peano Arithmetic and based on the self-referential structure. Ours also involves a self-referential structure, but our undecidability arises from some discord in interpersonal thinking in the self-referential environment.

In our problem, the minds of two players are described separately in the logic IR^2, but they have no effective differences before the description of game payoffs is given, since they have full logical abilities and the same prediction/decision criteria. Only the infinite regress of game payoffs $\mathbf{Ir}_i(g_i; g_j)$ in $\mathbf{\Delta}_i$ differentiates the two players. This difference is the source for the undecidability (4). A comparison with Gödel's theorem will be discussed in Section 6.

Other epistemic axioms: To accommodate the considerations of (3) and (4), as stated, we adopt the fixed-point extension IR^2 of the epistemic logic KD^2. In fact, both results of (3) and (4) hold for a stronger system than IR^2, for example, in those with Axioms T, 4, and 5. The reason for the choice of KD^2 is to keep a clear-cut structure of nested belief hierarchy of beliefs and to keep separate subjectivity of individual minds. Nevertheless, in particular, the addition of Axiom T is relevant and will be discussed in a few places.

Extensions to the n-person case: In the present paper, we confine ourselves to the 2-person case both for the logic and game theory. For n-person case ($n \geq 3$), we would meet new problems in both epistemic logic and game theory. We will discuss those extensions in separate papers.

The format of the paper is as follows: Section 2 formulates the logic IR^2. Section 3 gives various game theoretical concepts. Section 4 gives three axioms for prediction/decision making, and the characterization theorem for an interchangeable game. Section 5 presents the undecidability result for an uninterchangeable game, and the no-formula theorem. Section 6 gives discussions on our undecidability relative to Gödel's incompleteness theorem.

2 The Infinite Regress Logic IR^2

We use an fixed-point extension IR^2 of the epistemic logic KD^2, in order to capture an infinite regress arising in prediction/decision making in a game with two players. We formulate the logic IR^2 in Sections 2.1, 2.2, and give its semantics in Section 2.3.

2.1 Language

Let S_i be a nonempty finite *strategy* set for player $i = 1, 2$. We adopt the atomic formulae:

atomic preference formulae: $\mathrm{Pr}_i(s;t)$ for $i = 1, 2$ and $s, t \in S = S_1 \times S_2$;

atomic decision formulae: $\mathrm{I}_i(s_i)$ for $s_i \in S_i$, $i = 1, 2$.

The atomic formula $\mathrm{Pr}_i(\cdot;\cdot)$ expresses the preference relation of player i; $\mathrm{Pr}_i(s;t)$ means that player i *weakly prefers* the strategy pair $s = (s_1, s_2)$ to the pair $t = (t_1, t_2)$. The atomic formula $\mathrm{I}_i(s_i)$ expresses the idea that, from player i's perspective, s_i is a *possible final decision* for him.

Now we proceed to have logical connectives and epistemic operators:

logical connective symbols: $\neg$ (not), $\supset$ (imply), $\wedge$ (and), $\vee$ (or);[3]

unary belief operators: $\mathbf{B}_1(\cdot)$, $\mathbf{B}_2(\cdot)$;

binary infinite-regress operators: $\mathbf{Ir}_1(\cdot,\cdot)$, $\mathbf{Ir}_2(\cdot,\cdot)$;

parentheses: (,).

We use a pair of formulae, (A_1, A_2), as arguments of the binary operators $\mathbf{Ir}_1(\cdot,\cdot)$ and $\mathbf{Ir}_2(\cdot,\cdot)$, and the intended meaning of the formula $\mathbf{Ir}_i(A_1, A_2)$ is that player i's subjective belief of the infinite regress of beliefs about A_i and A_j. We stipulate that j refers to the other player than i. We write $\mathbf{Ir}_i(A_1, A_2)$ also as $\mathbf{Ir}_i(A_i; A_j)$ and sometimes $\mathbf{Ir}_i[A_i; A_j]$.

We define the sets of *formulae*, denoted by $\mathcal{P}$, by the following induction:

(o) all atomic formulae are formulae;

(i) if A, B are formulae, then so are $(A \supset B)$, $(\neg A)$, $\mathbf{B}_i(A)$ for $i = 1, 2$;

(ii) if $\mathbf{A} = (A_1, A_2)$ is a pair of formulae, then $\mathbf{Ir}_i(\mathbf{A})$ is also a formula;

(iii) if Φ is a finite (nonempty) set of formulae, then $(\wedge\Phi)$ and $(\vee\Phi)$ are formulae[4].

We say that a formula A is *non-epistemic* iff $\mathbf{B}_i(\cdot)$ or $\mathbf{Ir}_i(\cdot,\cdot)$ does not occur in A for $i = 1, 2$. We say that A_i is a *game formula for i* iff it contains atomic formulae of the form $\mathrm{Pr}_i(\cdot;\cdot)$ only, that is, no occurrences of $\mathrm{Pr}_j(\cdot;\cdot)$, $\mathrm{I}_i(\cdot)$, or $\mathrm{I}_j(\cdot)$; and that A is a *game formula* iff the atomic formulae occurring in A are of the form $\mathrm{Pr}_1(\cdot;\cdot)$ or $\mathrm{Pr}_2(\cdot;\cdot)$. A game formula expresses a reality of the target situation together with, potentially, beliefs about them. The atomic decision formulae $\mathrm{I}_i(s_i)$'s are used to describe a player's thinking about prediction/decision making.

We write $\wedge\{A, B\}$, $\wedge\{A, B, C\}$ as $A \wedge B$, $A \wedge B \wedge C$, etc., and $(A \supset B) \wedge (B \supset A)$ as $A \equiv B$. We abbreviate parentheses or use different ones such as $[,]$ when no confusions are expected.

[3] Since we adopt classical logic as the base logic, we can abbreviate some of those connectives. Since, however, our aim is to study logical inference for decision making rather than semantic contents, we use a full system.

[4] We presume the identity of "finite sets" in our language.

2.2 Proof theory of $\mathbf{IR}^2$

The base logic of IR^2 is classical logic, formulated by five axiom (schemata) and three inference rules: for all formulae A, B, C, and finite nonempty sets Φ of formulae,

L1 $A \supset (B \supset A)$;

L2 $(A \supset (B \supset C)) \supset ((A \supset B) \supset (A \supset C))$;

L3 $(\neg A \supset \neg B) \supset ((\neg A \supset B) \supset A)$;

L4 $\wedge\Phi \supset A$, where $A \in \Phi$;

L5 $A \supset \vee\Phi$, where $A \in \Phi$;

$$\frac{A \supset B \quad A}{B}\ \mathbf{MP} \qquad \frac{\{A \supset B : B \in \Phi\}}{A \supset \wedge\Phi}\ \wedge\text{-}\mathbf{rule} \qquad \frac{\{B \supset A : B \in \Phi\}}{\vee\Phi \supset A}\ \vee\text{-}\mathbf{rule}.$$

Now, we add two epistemic axioms and one inference rule for the belief operators $\mathbf{B}_i(\cdot)$: for all formulae A, B, and for $i = 1, 2$,

K $\mathbf{B}_i(A \supset B) \supset (\mathbf{B}_i(A) \supset \mathbf{B}_i(B))$;

D $\neg\mathbf{B}_i(\neg A \wedge A)$;

Necessitation $\dfrac{A}{\mathbf{B}_i(A)}$.

Those axioms and inference rules constitute the epistemic logic KD^2.

For the infinite regress operators $\mathbf{Ir}_i(\cdot,\cdot)$, we add one axiom and one inference rule: For $i = 1, 2$, and $\mathbf{A} = (A_1, A_2)$, $\mathbf{D} = (D_1, D_2)$ two pairs of formulae,

IRA$_i$ $\mathbf{Ir}_i(\mathbf{A}) \supset \mathbf{B}_i(A_i) \wedge \mathbf{B}_i\mathbf{B}_j(A_j) \wedge \mathbf{B}_i\mathbf{B}_j(\mathbf{Ir}_i(\mathbf{A}))$;

IRI$_i$ $\dfrac{D_i \supset \mathbf{B}_i(A_i) \wedge \mathbf{B}_i\mathbf{B}_j(A_j) \wedge \mathbf{B}_i\mathbf{B}_j(D_i)}{D_i \supset \mathbf{Ir}_i(\mathbf{A})}$.

The logic IR^2 is defined by adding IRA_i and IRI_i, $i = 1, 2$, to KD^2.

Axiom IRA_i has a fixed-point structure in the sense that $\mathbf{B}_i\mathbf{B}_j(\mathbf{Ir}_i(\mathbf{A}))$ appears as an implication of $\mathbf{Ir}_i(\mathbf{A})$. Replacing $\mathbf{Ir}_i(\mathbf{A})$ in $\mathbf{B}_i\mathbf{B}_j(\mathbf{Ir}_i(\mathbf{A}))$ with its implication $\mathbf{B}_i(A_i) \wedge \mathbf{B}_i\mathbf{B}_j(A_j)$ (formally with K and Nec), $\mathbf{Ir}_i(\mathbf{A})$ implies the following infinite regress of beliefs:

$$\{\mathbf{B}_i(A_i), \mathbf{B}_i\mathbf{B}_j(A_j), \mathbf{B}_i\mathbf{B}_j\mathbf{B}_i(A_i), ...\}. \tag{5}$$

Rule IRI_i states that $\mathbf{Ir}_i(\mathbf{A})$ is the logically weakest formula satisfying the property described in IRA_i, that is, if D_i enjoys it, then D_i implies $\mathbf{Ir}_i(\mathbf{A})$. Our completeness-soundness theorem (Theorem 2.1) shows that $\mathbf{Ir}_i(\mathbf{A})$ captures faithfully the set of (5).

A *proof* $P = \langle X, <; \psi\rangle$ consists of a finite tree $\langle X, <\rangle$ and a function $\psi : X \to \mathcal{P}$ with the following requirements:

P1 for each node $x \in X$, $\psi(x)$ is a formula attached to x;

P2 for each leaf x in $\langle X, <\rangle$, $\psi(x)$ is an instance of the axiom schemata;

P3 for each non-leaf x in $\langle X, <\rangle$,

$$\frac{\{\psi(y) :\ y \text{ is an immediate predecessor of } x\}}{\psi(x)}$$

is an instance of the above five inference rules.

We call P *a proof of* A iff $\psi(x_0) = A$, where x_0 is the root of $\langle X, < \rangle$. We say that A is *provable*, denoted by $\vdash A$, iff there is a proof of A. For a set of formulae Γ, we write $\Gamma \vdash A$ iff $\vdash A$ or there is a finite nonempty subset Φ of Γ such that $\vdash \wedge\Phi \supset A$. This treatment of non-logical assumptions is crucial in our study[5].

The following are basic to classical logic and/or KD^2. We use them without referring.

Lemma 2.1. *Let $A \in \mathcal{P}$, Φ a finite set of formulae, and $i = 1, 2$. Then,* ***(1)*** $\vdash A \supset B$ *and* $\vdash B \supset C$ *imply* $\vdash A \supset C$; ***(2)*** $\vdash (A \wedge B \supset C) \equiv (A \supset (B \supset C))$; ***(3)*** $\vdash \mathbf{B}_i(\neg A) \supset \neg\mathbf{B}_i(A)$; ***(4)*** $\vdash \vee\mathbf{B}_i(\Phi) \supset \mathbf{B}_i(\vee\Phi)$; ***(5)*** $\vdash \mathbf{B}_i(\wedge\Phi) \equiv \wedge\mathbf{B}_i(\Phi)$.

From Axiom IRA$_i$ and Rule IRI$_i$ $(i = 1, 2)$, the operators $\mathbf{Ir}_i(\cdot, \cdot)$, $i = 1, 2$ may appear to be independent of one another. However, the two operators are interdependent:

Lemma 2.2. ***(Epistemic content)*** *Let $\mathbf{A} = (A_1, A_2)$ be a pair of formulae. Then,* $\vdash \mathbf{Ir}_i(\mathbf{A}) \equiv \mathbf{B}_i(A_i \wedge \mathbf{Ir}_j(\mathbf{A}))$ *for $i = 1, 2$.*

Proof. First, we show $\vdash \mathbf{B}_i(A_i \wedge \mathbf{Ir}_j(\mathbf{A})) \supset \mathbf{Ir}_i(\mathbf{A})$. Let $D_i = \mathbf{B}_i(A_i) \wedge \mathbf{B}_i(\mathbf{Ir}_j(\mathbf{A}))$ for $i = 1, 2$. By IRA$_j$ (and, Nec, K), we have $\vdash D_i \supset \mathbf{B}_i(A_i) \wedge \mathbf{B}_i\mathbf{B}_j(A_j) \wedge \mathbf{B}_i\mathbf{B}_j\mathbf{B}_i(A_i) \wedge \mathbf{B}_i\mathbf{B}_j\mathbf{B}_i(\mathbf{Ir}_j(\mathbf{A}))$. Since the last two conjuncts are equivalent to $\mathbf{B}_i\mathbf{B}_j(D_i)$, we have $\vdash D_i \supset \mathbf{B}_i(A_i) \wedge \mathbf{B}_i\mathbf{B}_j(A_j) \wedge \mathbf{B}_i\mathbf{B}_j(D_i)$. Using IRI$_i$, we have $\vdash \mathbf{B}_i(A_i) \wedge \mathbf{B}_i(\mathbf{Ir}_j(\mathbf{A})) \supset \mathbf{Ir}_i(\mathbf{A})$.

The above conclusion for j implies $\vdash \mathbf{B}_i(D_j) \supset \mathbf{B}_i(\mathbf{Ir}_j(\mathbf{A}))$. Hence, we have $\vdash \mathbf{B}_i(A_i) \wedge \mathbf{B}_i(D_j) \supset \mathbf{B}_i(A_i) \wedge \mathbf{B}_i(\mathbf{Ir}_j(\mathbf{A}))$. Since $\vdash \mathbf{Ir}_i(\mathbf{A}) \supset \mathbf{B}_i(A_i) \wedge \mathbf{B}_i(D_j)$ by IRA$_i$, we have $\vdash \mathbf{Ir}_i(\mathbf{A}) \supset \mathbf{B}_i(A_i) \wedge \mathbf{B}_i(\mathbf{Ir}_j(\mathbf{A}))$.$\blacksquare$

This lemma enables us to talk about the *epistemic content* of $\mathbf{Ir}_i(\mathbf{A})$;

$$\mathbf{Ir}_i^o(\mathbf{A}) := A_i \wedge \mathbf{Ir}_j(\mathbf{A}), \tag{6}$$

which plays a crucial role in our consideration of prediction/decision making.

Lemma 2.3. ***(Basic properties for*** $Ir_i(\cdot;\cdot)$***)*** *Let $\mathbf{A} = (A_1, A_2)$ and $\mathbf{C} = (C_1, C_2)$ be two pairs of formulae in $\mathcal{P}$ and $i = 1, 2$.*

(1) *If* $\vdash \mathbf{Ir}_k(\mathbf{A}) \supset \mathbf{B}_k(C_k)$ *for $k = 1, 2$, then* $\vdash \mathbf{Ir}_i(\mathbf{A}) \supset \mathbf{Ir}_i(\mathbf{C})$. *In particular, if* $\vdash C_i$ *for $i = 1, 2$, then* $\vdash \mathbf{Ir}_i(\mathbf{C})$.

(2) $\vdash \mathbf{Ir}_i(\mathbf{A}) \supset \mathbf{Ir}_i(\mathbf{Ir}_1^o(\mathbf{A}), \mathbf{Ir}_2^o(\mathbf{A}))$;

(3) $\vdash \mathbf{Ir}_i(A_1 \wedge C_1, A_2 \wedge C_2) \equiv \mathbf{Ir}_i(\mathbf{A}) \wedge \mathbf{Ir}_i(\mathbf{C})$;

(4) $\vdash \mathbf{Ir}_i(A_1 \supset C_1, A_2 \supset C_2) \supset (\mathbf{Ir}_i(\mathbf{A}) \supset \mathbf{Ir}_i(\mathbf{C}))$:

(5) $\vdash \mathbf{Ir}_i(\neg A_i; A_j) \supset \neg\mathbf{Ir}_i(A_i; A_j)$; $\vdash \mathbf{Ir}_i(A_i; \neg A_j) \supset \neg\mathbf{Ir}_i(A_i; A_j)$ *and* $\vdash \mathbf{Ir}_i(\neg A_i; \neg A_j) \supset \neg\mathbf{Ir}_i(A_i; A_j)$.

Proof. (1): Let $\vdash \mathbf{Ir}_k(\mathbf{A}) \supset \mathbf{B}_k(C_k)$ for $k = 1, 2$. We show $\vdash \mathbf{Ir}_i(\mathbf{A}) \supset \mathbf{B}_i(C_i) \wedge \mathbf{B}_i\mathbf{B}_j(C_j) \wedge \mathbf{B}_i\mathbf{B}_j(\mathbf{Ir}_i(\mathbf{A}))$. Once this is shown, we have, by IRI$_i$, $\vdash \mathbf{Ir}_i(\mathbf{A}) \supset \mathbf{Ir}_i(\mathbf{C})$. First, $\vdash \mathbf{B}_i(\mathbf{Ir}_j(\mathbf{A})) \supset$

[5] Since the deduction theorem (cf., Mendelson [16]) does not hold in epistemic logic, the introduction of non-logical axioms differs from in classical logic. We adopt the classical manner.

$\mathbf{B}_i\mathbf{B}_j(C_j)$ by Nec and K. By Lemma 2.2, we have $\vdash \mathbf{Ir}_i(\mathbf{A}) \supset \mathbf{B}_i\mathbf{B}_j(C_j)$. By IRA$_i$, we have $\vdash \mathbf{Ir}_i(\mathbf{A}) \supset \mathbf{B}_i\mathbf{B}_j(\mathbf{Ir}_i(\mathbf{A}))$. Thus, by $\wedge$-rule, we have the target.

The other claims (2)-(4) follow (1). Here, we show (3). Since $\vdash \mathbf{Ir}_k(A_1 \wedge C_1, A_2 \wedge C_2) \supset \mathbf{B}_k(A_k)$ for $k = 1, 2$, we have, by (1), $\vdash \mathbf{Ir}_k(A_1 \wedge C_1, A_2 \wedge C_2) \supset \mathbf{Ir}_i(\mathbf{A})$. Similarly, $\vdash \mathbf{Ir}_k(A_1 \wedge C_1, A_2 \wedge C_2) \supset \mathbf{Ir}_i(\mathbf{C})$. Hence, we have the one direction. Consider the converse. We have $\vdash \mathbf{Ir}_k(\mathbf{A}) \wedge \mathbf{Ir}_k(\mathbf{C}) \supset \mathbf{B}_k(A_k \wedge C_k)$ for $k = 1, 2$. We have $\vdash \mathbf{Ir}_i(\mathbf{A}) \wedge \mathbf{Ir}_i(\mathbf{C}) \supset \mathbf{B}_i\mathbf{B}_j(A_j \wedge C_j)$, and $\vdash \mathbf{Ir}_i(\mathbf{A}) \wedge \mathbf{Ir}_i(\mathbf{C}) \supset \mathbf{B}_i\mathbf{B}_j(\mathbf{Ir}_i(\mathbf{A}) \wedge \mathbf{Ir}_i(\mathbf{C}))$. Then, by IRI$_i$, $\vdash \mathbf{Ir}_i(\mathbf{A}) \wedge \mathbf{Ir}_i(\mathbf{C}) \supset \mathbf{Ir}_i(A_1 \wedge C_1, A_1 \wedge C_2)$.

(5): Consider only the first one. Since $\vdash \mathbf{Ir}_i(\neg A_i; A_j) \supset \mathbf{B}(\neg A_i)$, we have $\vdash \mathbf{Ir}_i(\neg A_i; A_j) \supset \neg\mathbf{B}(A_i)$. Then, using the contrapositive of IRA$_i$, i.e., $\vdash \neg[\mathbf{B}_i(A_i) \wedge \mathbf{B}_i\mathbf{B}_j(A_j) \wedge \mathbf{B}_i\mathbf{B}_j(\mathbf{Ir}_i(\mathbf{A}))] \supset \neg\mathbf{Ir}_i(\mathbf{A})$, we have $\vdash \mathbf{Ir}_i(\neg A_i; A_j) \supset \neg\mathbf{Ir}_i(\mathbf{A})$.■

The following statements for $\mathbf{Ir}_i^o(\cdot;\cdot)$ correspond to IRA$_i$ and IRI$_i$ for $\mathbf{Ir}_i(\cdot;\cdot)$.

Lemma 2.4. ***(Admissible formulae and inference)*** *Let $\mathbf{A} = (A_i; A_j)$ and D_i be any formulae. Then,*

(IRA$_i^o$) $\vdash \mathbf{Ir}_i^o(\mathbf{A}) \supset A_i \wedge \mathbf{B}_j(A_j) \wedge \mathbf{B}_j\mathbf{B}_i(\mathbf{Ir}_i^o(\mathbf{A}))$;

(IRI$_i^o$) *If* $\vdash D_i \supset A_i \wedge \mathbf{B}_j(A_j) \wedge \mathbf{B}_j\mathbf{B}_i(D_i)$, *then* $\vdash D_i \supset \mathbf{Ir}_i^o(A_i; A_j)$.

Proof. (IRA$_i^o$): By (6), $\vdash \mathbf{Ir}_i^o(\mathbf{A}) \supset A_i \wedge \mathbf{Ir}_j(\mathbf{A})$. By Lemma 2.2 for j, we have $\vdash \mathbf{Ir}_i^o(\mathbf{A}) \supset A_i \wedge \mathbf{B}_j(A_j) \wedge \mathbf{B}_j(\mathbf{Ir}_i(\mathbf{A}))$, which is (1).

(IRI$_i^o$): Suppose $\vdash D_i \supset A_i \wedge \mathbf{B}_j(A_j) \wedge \mathbf{B}_j\mathbf{B}_i(D_i)$. Since $\vdash D_i \supset \mathbf{B}_j\mathbf{B}_i(D_i)$ and $\vdash D_i \supset A_i$, we have $\vdash D_i \supset \mathbf{B}_j\mathbf{B}_i(A_i)$. Thus, $\vdash D_i \supset \mathbf{B}_j(A_j) \wedge \mathbf{B}_j\mathbf{B}_i(A_i) \wedge \mathbf{B}_j\mathbf{B}_i(D_i)$. By IRI$_i$, we have $\vdash D_i \supset \mathbf{Ir}_j(A_i; A_j)$. Thus, $\vdash D_i \supset A_i \wedge \mathbf{Ir}_j(A_i; A_j)$, which is $\vdash D_i \supset \mathbf{Ir}_i^o(A_i; A_j)$.■

The main undecidability result of the paper holds in a stronger system than IR^2, such as that obtained from IR^2 by adding Axiom T (truthfulness): $\mathbf{B}_i(A) \supset A$; Axiom 4 (positive introspection): $\mathbf{B}_i(A) \supset \mathbf{B}_i\mathbf{B}_i(A)$; and Axiom 5 (negative introspection): $\neg\mathbf{B}_i(A) \supset \mathbf{B}_i(\neg\mathbf{B}_i(A))$. The reason for our choice of IR^2 is to have a clear-cut description of each player's logical inference. This is stated by Lemma 2.5 (change of scopes), which is specific to IR^2. Nevertheless, Axiom T helps us understand the fixed-point formula $\mathbf{Ir}_i(A_1, A_2)$.

Now, let us see the common knowledge logic CKL (cf., Fagin *et al.* [5] and Meyer-van der Hoek [15]). The logic CKL uses only one operator, $\mathbf{C}(\cdot)$, and adds the following axiom and rule to KD^2:

CKA: $\mathbf{C}(A) \supset A \wedge \mathbf{B}_1(\mathbf{C}(A)) \wedge \mathbf{B}_2(\mathbf{C}(A))$;

CKI: $\dfrac{D \supset A \wedge \mathbf{B}_1(D) \wedge \mathbf{B}_2(D)}{D \supset \mathbf{C}(A)}$.

Axiom CKA and Rule CKI are interpreted as meaning that $\mathbf{C}(A)$ describes the common knowledge of A from the outside observer's perspective; on the other hand, $\mathbf{Ir}_i(\mathbf{A})$ describes player i's subjective beliefs from his perspective. This difference is reflected by the counterpart of (5) in CKL, i.e., $\mathbf{C}(A)$ captures the entire set:

$$\{A, \mathbf{B}_1(A), \mathbf{B}_2(A), \mathbf{B}_1\mathbf{B}_2(A), \mathbf{B}_2\mathbf{B}_1(A), \mathbf{B}_1\mathbf{B}_2\mathbf{B}_2(A), ...\}. \tag{7}$$

This set of formulae having all finite sequences of $\mathbf{B}_2\mathbf{B}_1...$ including the repetitive ones such as $\mathbf{B}_1\mathbf{B}_2\mathbf{B}_2$, while each in (5) has the outer $\mathbf{B}_i(\cdot)$ and all $\mathbf{B}_i\mathbf{B}_j...$ are alternating.

Let us look at IR^2 with Axiom T. The logical system obtained from IR^2 by adding Axiom T is denoted by $\mathrm{IR}^2(\mathrm{T})$. Since $\vdash \mathbf{Ir}_1(A_1, A_2) \equiv \mathbf{Ir}_2(A_1, A_2)$ in $\mathrm{IR}^2(\mathrm{T})$ by Lemma 2.2, we can denote $\mathbf{Ir}_i(A_1, A_2)$ by $\mathbf{C}^*(A_1 \wedge A_2)$. Then, in $\mathrm{IR}^2(\mathrm{T})$, we have, for any formulae A_1, A_2 and D,

cka: $\vdash \mathbf{C}^*(A_1 \wedge A_2) \supset (A_1 \wedge A_2) \wedge \mathbf{B}_1\mathbf{C}^*(A_1 \wedge A_2) \wedge \mathbf{B}_2\mathbf{C}^*(A_1 \wedge A_2)$;

cki: if $\vdash D \supset (A_1 \wedge A_2) \wedge \mathbf{B}_1(D) \wedge \mathbf{B}_2(D)$, then $\vdash D \supset \mathbf{C}^*(A_1 \wedge A_2)$.

These mean that in $\mathrm{IR}^2(\mathrm{T})$, CKA and CKI are derived formulae and admissible rule for $\mathbf{C}^*(A_1 \wedge A_2)$. Thus, $\mathbf{C}^*(A_1 \wedge A_2)$ $(= \mathbf{Ir}_i(A_1, A_2))$ means the common knowledge of $A_1 \wedge A_2$.

We will use the *belief eraser* ε_0 : the formula $\varepsilon_0(A) \in \mathcal{P}_N$ is obtained from $A \in \mathcal{P}$ by eliminating all occurrences of $\mathbf{B}_1(\cdot), \mathbf{B}_2(\cdot)$ and replacing $\mathbf{Ir}_i(A_1, A_2)$ by $\varepsilon_0(A_1) \wedge \varepsilon_0(A_2)$. Then, we have

$$\vdash A \text{ implies } \vdash_0 \varepsilon_0(A), \tag{8}$$

where $\vdash_0$ is the provability relation of classical logic in $\mathcal{P}_N$. This is proved by induction on a proof of A from leaves (Kaneko-Nagashima [11]).

2.3 Kripke semantics and the soundness/completeness of $\mathbf{IR}^2$

Here, we report the soundness/completeness for IR^2 with respect to the Kripke semantics. We use the soundness part for the main undecidability result.

A Kripke frame $\langle W; R_1, R_2\rangle$ consists of a nonempty set W of possible worlds and an accessibility relation R_i for player $i = 1, 2$. We say that a frame $\langle W; R_1, R_2\rangle$ is *serial* iff for $i = 1, 2$ and for all $w \in W$, wR_iu for some $u \in W$. A *truth assignment* τ is a function from $W \times AF$ to $\{\top, \bot\}$, where AF is the set of atomic formulae. A pair $M = (\langle W; R_1, R_2\rangle, \tau)$ is called a *model.* When $\langle W; R_1, R_2\rangle$ is serial, we say that M is a serial model.

We say that $\langle (w_0, i_0), ..., (w_\nu, i_\nu), w_{\nu+1}\rangle$ $(\nu \geq 0)$ is an *alternating sequence* from (w_0, i_0) iff $i_{k-1} \neq i_{i_k}$ for $k = 1, ..., \nu$ and $w_{k-1}R_{i_{k-1}}w_k$ for $k = 1, ..., \nu + 1$. The alternating structure corresponds to the set given by (5). This is used for evaluating the truth values of formulae $\mathbf{Ir}_i(A_1, A_2)$, $i = 1, 2$.

The valuation in (M, w), denoted by $(M, w) \models$, is defined over $\mathcal{P}$ by induction on the length of a formula as follows:

V0 for any $A \in AF$, $(M, w) \models A \Longleftrightarrow \tau(w, A) = \top$;

V1 $(M, w) \models \neg A \Longleftrightarrow (M, w) \not\models A$;

V2 $(M, w) \models A \supset B \Longleftrightarrow$ $(M, w) \not\models A$ or $(M, w) \models B$;

V3 $(M, w) \models \wedge\Phi \Longleftrightarrow (M, w) \models A$ for all $A \in \Phi$;

V4 $(M, w) \models \vee\Phi \Longleftrightarrow (M, w) \models A$ for some $A \in \Phi$;

V5 $(M, w) \models \mathbf{B}_i(A) \Longleftrightarrow (M, v) \models A$ for all v with wR_iv;

V6 $(M, w) \models \mathbf{Ir}_i(A_1, A_2) \Longleftrightarrow (M, w_{\nu+1}) \models A_{i_\nu}$ for any alternating sequence $\langle (w_0, i_0), ...,$ $(w_\nu, i_\nu), w_{\nu+1}\rangle$ with $(w_0, i_0) = (w, i)$.

The steps other than V6 are standard. V6 is similar to the valuation for the common knowledge operator in CKL; the only difference is to use alternating reachability for two formulae, instead of simple rearchability (cf., Fagin *et al.* [5], Meyer-van der Hoek [15]).

We have the following soundness/completeness theorem.

Theorem 2.1. ***(Soundness and Completeness)*** *Let $A \in \mathcal{P}$. Then, $\vdash A$ in IR^2 if and only if $(M, w) \models A$ for all serial models $M = (\langle W; R_1, R_2 \rangle, \tau)$ and any $w \in W$.*

Soundness (only-if) will be used to prove our undecidability result (Theorem 5.1). It is proved as follows: Let $P = (X, <; \psi)$ be a proof of A. Then, by induction on the tree structure of $(X, <)$ from its leaves, we show that for any $x \in X$, $\vdash \psi(x)$ implies $\models \psi(x)$. The two new steps are : (1) $\models C$ for any instance C of IRA_i; and (2) the validity relation $\models$ satisfies IRA_i. Both steps follow V6. The proof of completeness is given in Hu-Kaneko [8].

Theorem 2.1 shows that our fixed-point operator $\mathbf{Ir}_i(\mathbf{A})$ faithfully captures the set of (5). The alternating structure in the semantics implies that if $\mathbf{Ir}_i(\mathbf{A})$ holds at a world w and if wR_iu, then A_i and $\mathbf{Ir}_j(\mathbf{A})$ hold at world u, which corresponds to Lemma 2.2, further if uR_iv, then $\mathbf{Ir}_i(\mathbf{A})$ holds at world v, which corresponds to IRA_i. These reflect the self-referential structure shared by $\mathbf{Ir}_i(\mathbf{A})$ and $\mathbf{Ir}_j(\mathbf{A})$.

The following lemma requires the logic IR^2 with its base logic KD^2, which is proved by both soundness and completeness of Theorem 2.1. The lemma does not hold for IR^2 with any addition of Axioms T, 4 and 5; counter examples are given in Hu-Kaneko [8].

Lemma 2.5. ***(Change of Scopes)*** *(1):* $\mathbf{B}_i(\Gamma_i^o) \vdash \mathbf{B}_i(A) \Longleftrightarrow \Gamma_i^o \vdash A$;

(2): $\mathbf{B}_i(\Gamma_i^o) \vdash \neg\mathbf{B}_i(A) \Longleftrightarrow \mathbf{B}_i(\Gamma_i^o) \vdash \mathbf{B}_i(\neg A)$.

In our application, $\mathbf{Ir}_i(A_1, A_2)$ is used as a premise of a statement of the form $\mathbf{Ir}_i(A_1, A_2) \vdash \mathbf{B}_i(C)$. By Lemmas 2.2 and 2.5, this is equivalent to $\mathbf{Ir}_i^o(A_1, A_2) \vdash C$. This is interpreted as meaning that $\mathbf{Ir}_i^o(A_1, A_2) \vdash C$ is obtained in the mind of player i.

3 Game Theoretic Concepts

First, we give a few game theoretic concepts relevant for our discussions. Then, we formulate them in the language of the logic IR^2. We also mention some decidability (completeness) for comparisons with the main undecidability result.

3.1 Preliminary definitions

Let $G = (\{1, 2\}, \{S_1, S_2\}, \{h_1, h_2\})$ be a finite 2-person game, where $\{1, 2\}$ is the set of players, $S = S_1 \times S_2$ is the set of *strategy pairs*, and $h_i : S \to \mathbb{R}$ is the payoff function for player $i = 1, 2$. We write $(s_i; s_j)$ for $s = (s_1, s_2) \in S$. A strategy s_i for player i is a *best-response* against s_j iff $h_i(s_i; s_j) \geq h_i(t_i; s_j)$ for all $t_i \in S_i$. A strategy pair $s = (s_i; s_j)$ is a *Nash equilibrium* in G iff s_i is a best response against s_j for $i = 1, 2$. We denote $E(G)$ the set of all Nash equilibria in G. The set $E(G)$ may be empty. We say that s_i is a *Nash strategy* iff $(s_i; s_j)$ is a Nash equilibrium for some $s_j \in S_j$. The game of Table 1.1 has a unique Nash equilibrium, and Table 1.2 have two, indicated by the superscript NE. Table 1.3 has no Nash equilibria.

A subset E of S is *interchangeable* (Nash [17]) iff

$$\text{for all } s, s' \in E,\ (s_i; s_j') \in E \text{ for } i = 1, 2. \tag{9}$$

This is equivalent to $E = E_1 \times E_2$, where $E_i = \{s_i \in S_i : (s_i; s_j) \in E$ for some $s_j\}$, $i = 1, 2$. Let $\mathbf{E} = \{E : E \subseteq E(G)$ and E satisfies (9)$\}$. A *nonempty* subset E of S is the *(Nash) solution* iff E is the greatest set in $\mathbf{E}$, i.e., $E' \subseteq E$ for any $E' \in \mathbf{E}$. The Nash solution, when it exists, is unique and coincides with $E(G)$. The game G is *solvable* iff G has the Nash solution; otherwise, it is *unsolvable*. A nonempty set $F \subseteq S$ is a *subsolution* iff F is a maximal set in $\mathbf{E}$, i.e., there is no $E' \in \mathbf{E}$ such that $F \subsetneq E'$. When G has a unique subsolution, it is the solution $E(G)$. Table 1.1 is solvable with the solution $\{(\mathbf{s}_{12}, \mathbf{s}_{21})\}$. Table 1.2 is unsolvable, and has two subsolutions: $\{(\mathbf{s}_{11}, \mathbf{s}_{21})\}$ and $\{(\mathbf{s}_{12}, \mathbf{s}_{22})\}$. Table 1.3 has no subsolution.

Nash [17] assumed the mixed strategies, and proved the existence of a Nash equilibrium. Here, some games have no Nash equilibria. For our considerations, it would be more convenient to separate the games with interchangeable $E(G)$ from the other games. Therefore, we call G an *interchangeable game* iff $E(G)$ is interchangeable. A game is interchangeable if and only if it has no subsolution or the unique subsolution; and G an *uninterchangeable game* iff it has multiple subsolutions.

Hu-Kaneko [7] derived the Nash (sub)solutions from the following decision criteria: Let E_i be a subset of S_i for $i = 1, 2$.

Na$_1$: for any $s_1 \in E_1$, s_1 is a best response against all $s_2 \in E_2$;

Na$_2$: for any $s_2 \in E_2$, s_2 is a best response against all $s_1 \in E_1$.

In Na$_i$, E_i describes the set of possible final decisions for player i, and E_j describes i's prediction about j's possible final decisions. Here i's prediction comes from his thinking about j's inference from j's basic beliefs. Specifically, player i assumes that j's basic beliefs consist of the decision criterion Na$_j$ and the game structure. In epistemic terms, when i makes his prediction based on E_j, elements in E_j occur in the scope of j's thinking, and this whole statement occurs in the scope of i's thinking. In the present language, we cannot make distinguish between i's and j's thinking, which are all interpretational. We will formalize this distinction in our logic IR^2.

The following proposition was proved in Hu-Kaneko [7].

Proposition 3.1. *Let $E(G) \neq \emptyset$, and E_i a nonempty subset of S_i for $i = 1, 2$.*

(1) Suppose that G is solvable. Then $E = E_1 \times E_2$ is the Nash solution of G if and only if (E_1, E_2) is the greatest pair satisfying Na$_1$-Na$_2$.[6]

(2) Suppose that G is unsolvable. Then $E = E_1 \times E_2$ is a Nash subsolution if and only if (E_1, E_2) is a maximal pair satisfying Na$_1$-Na$_2$.

These two cases correspond basically to the decidability and undecidability results to be discussed in the subsequent sections. Here, we avoided unnecessary complication for the case of $E(G) = \emptyset$. In the subsequent sections, we treat that case, too.

3.2 Game formulae in IR^2 and some decidabilities

For the description of a game $G = \langle\{1, 2\}, \{S_1, S_2\}, \{h_1, h_2\}\rangle$ in the language of IR^2, it suffices to express the payoff functions h_1 and h_2, because the players and strategies are already included in the language. The payoff functions are expressed in terms of atomic preference formulae as

[6] The "greatest" and "maximal" are relative to the componentwise set-inclusions.

follows:

$$g_i = \wedge\left[\{\mathrm{Pr}_i(s;t) : h_i(s) \geq h_i(t)\} \cup \{\neg\mathrm{Pr}_i(s;t) : h_i(s) < h_i(t)\}\right]. \tag{10}$$

We call g_i the *formalized payoffs* associated with h_i for $i = 1, 2$. Since the latter part consists negative preferences, it holds that for all $s, t \in S$, $g_i \vdash \mathrm{Pr}_i(s;t)$ or $g_i \vdash \neg\mathrm{Pr}_i(s;t)$, i.e., g_i gives a complete preference relation.

Consistency of $g_1 \wedge g_2$ can be shown by constructing a truth assignment. The infinite regress $\mathbf{Ir}_i(g_1, g_2)$ is consistent in IR^2 is obtained by applying the belief eraser ε_0 : Suppose that $\mathbf{Ir}_i(g_1, g_2) \vdash \neg A \wedge A$ for some nonepistemic formula A. Applying ε_0 to this, we have $g_1 \wedge g_2 \vdash_0 \neg A \wedge A$ by (8), which is impossible because of the consistency of $g_1 \wedge g_2$. Consistency of $\mathbf{Ir}_i^o(g_1, g_2)$ in IR^2 follows, too. These are listed for reference.

$$\mathbf{Ir}_i(g_1, g_2) \text{ and } \mathbf{Ir}_i^o(g_1, g_2) \text{ are consistent in } \mathrm{IR}^2. \tag{11}$$

We formalize *best response* and *Nash equilibrium*: The statement "$s_i \in S_i$ is a best response to $s_j \in S_j$" is given as $\mathrm{bst}_i(s_i; s_j) := \wedge_{t_i \in S_i} \mathrm{Pr}_i(s_i, s_j; t_i, s_j)$. The statement "$s = (s_1, s_2) \in S$ is a Nash equilibrium" is given as $\mathrm{nash}(s) := \mathrm{bst}_1(s_1; s_2) \wedge \mathrm{bst}_2(s_2; s_1)$.

The formulae defined above are game formulae. The atomic formulae $\mathrm{I}_i(s_i)$ and $\mathrm{I}_j(s_j)$ are not included in them; they are used to describe prediction/decision making. Later, we will ask whether those are described directly by game formulae; this question is important in interpreting our undecidability as well as decidability.

We should assume that player i has enough beliefs, in order for the undecidability question to make sense. Undecidability could be an easy conclusion, if a belief set for player i has a weak content. As far as game formulae are concerned, the infinite regress of the formalized payoffs $\mathbf{Ir}_i(g_1, g_2)$ contains sufficient information to prove or to disprove them.

Lemma 3.1. *Let A_i be a nonepistemic game formula for $i = 1, 2$. Let G be a game and $\mathbf{g} = (g_1, g_2)$ the formalized payoffs. Then,*

(1) *$g_i \vdash A_i$ or $g_i \vdash \neg A_i$ for $i = 1, 2$;*

(2) *the following three are equivalent*

(a) *$\mathbf{Ir}_i(\mathbf{g}) \vdash \mathbf{Ir}_i(\mathbf{A})$ for $i = 1, 2$;* ***(b)*** *$\mathbf{Ir}_i^o(\mathbf{g}) \vdash \mathbf{Ir}_i^o(\mathbf{A})$ for $i = 1, 2$;* ***(c)*** *$g_i \vdash A_i$ for $i = 1, 2$.*

Proof. (1) Let $\mathrm{Pr}_i(s;t)$ be any atomic formula. Recall that $g_i \vdash \mathrm{Pr}_i(s;t)$ or $g_i \vdash \neg\mathrm{Pr}_i(s;t)$. We can extend this result to other nonepistemic game formulae for i by induction on their lengths.

(2) ($(c) \Longrightarrow (a) \Longrightarrow (b)$): Suppose that $g_i \vdash A_i$, i.e., $\vdash g_i \supset A_i$ for $i = 1, 2$. It follows from Lemma 2.3.(1) that $\vdash \mathbf{Ir}_i(g_1 \supset A_1, g_2 \supset A_2)$. By Lemma 2.3 (4), $\mathbf{Ir}_i(\mathbf{g}) \vdash \mathbf{Ir}_i(\mathbf{A})$ for $i = 1, 2$. Since $\vdash g_i \supset A_i$, we have $g_i \wedge \mathbf{Ir}_j(\mathbf{g}) \vdash A_i \wedge \mathbf{Ir}_j(\mathbf{A})$, which implies $\mathbf{Ir}_i^o(\mathbf{g}) \vdash \mathbf{Ir}_i^o(\mathbf{A})$.

($(b) \Longrightarrow (c)$): We show the contrapositive. Suppose that $g_1 \nvdash A_1$ or $g_2 \nvdash A_2$. By (1), $g_i \vdash \neg A_i$ or $g_j \vdash \neg A_j$ or both. We only consider the case where $g_i \vdash A_i$ and $g_j \vdash \neg A_j$. Using the same arguments as above, $\mathbf{Ir}_i^o(\mathbf{g}) \vdash \mathbf{Ir}_i^o(A_i; \neg A_j)$. By Lemma 2.4.(1), $\mathbf{Ir}_i^o(\mathbf{g}) \vdash \mathbf{B}_j(\neg A_j)$ and hence, $\mathbf{Ir}_i^o(\mathbf{g}) \vdash \neg\mathbf{B}_j(A_j)$. But by Lemma 2.4.(1), $\vdash \mathbf{Ir}_i^o(\mathbf{A}) \supset \mathbf{B}_j(A_j)$, and hence $\vdash \neg\mathbf{B}_j(A_j) \supset \neg\mathbf{Ir}_i^o(\mathbf{A})$. Therefore, $\mathbf{Ir}_i^o(\mathbf{g}) \vdash \neg\mathbf{Ir}_i^o(A_i; A_j)$. By (11), we have $\mathbf{Ir}_i^o(\mathbf{g}) \nvdash \mathbf{Ir}_i^o(A_i; A_j)$. In the other cases, we have similar arguments.■

Theorem 3.1 states that $\mathbf{Ir}_i(\mathbf{g})$ is enough for decidability as far as an infinite regress of nonepistemic game formulae concerned. It states this in terms of the epistemic content $\mathbf{Ir}_i^o(\cdot;\cdot)$ for coherency of the later aim.

Theorem 3.1. ***(Decidability for the infinite regress of game formulae)*** *Let G be a game and $\mathbf{g} = (g_1, g_2)$ the formalized payoffs. Let A_i be a nonepistemic game formula for $i = 1, 2$. Then, either $\mathbf{Ir}_i^o(\mathbf{g}) \vdash \mathbf{Ir}_i^o(\mathbf{A})$ or $\mathbf{Ir}_i^o(\mathbf{g}) \vdash \neg\mathbf{Ir}_i^o(\mathbf{A})$, which implies either $\mathbf{Ir}_i(\mathbf{g}) \vdash \mathbf{Ir}_i(\mathbf{A})$ or $\mathbf{Ir}_i(\mathbf{g}) \vdash \neg\mathbf{Ir}_i(\mathbf{A})$.*

Proof. Since $g_i \vdash A_i$ or $g_i \vdash \neg A_i$ for $i = 1, 2$, we should consider the four cases. Here, we consider only the case where $g_i \vdash \neg A_i$ for $i = 1, 2$. By (6), $\mathbf{Ir}_i^o(\mathbf{g}) \vdash \neg A_i$. Using the contrapositive of Lemma 2.4.(1), we have $\vdash \neg A_i \supset \neg\mathbf{Ir}_i^o(A_i; A_i)$. Thus, $\mathbf{Ir}_i^o(\mathbf{g}) \vdash \neg\mathbf{Ir}_i^o(A_i; A_i)$.■

Theorem 3.1 will be used for a positive result. We will discuss a negative result, too: For this purpose, we strengthen the logic to IR^2(T) by adding Axiom T. This theorem will be used for the no-formula theorem (Theorem 5.2).

Theorem 3.2. ***(Decidability for any game formula in IR^2(T))*** *Let G be a game and $\mathbf{g} = (g_1, g_2)$ the formalized payoffs. For any game formula A, either $\mathbf{Ir}_i(\mathbf{g}) \vdash A$ or $\mathbf{Ir}_i(\mathbf{g}) \vdash \neg A$ in IR^2(T).*

Proof. We prove the claim $\mathbf{Ir}_i^o(\mathbf{g}) \vdash A$ or $\mathbf{Ir}_i^o(\mathbf{g}) \vdash \neg A$ by induction on the length of A. This implies $\mathbf{Ir}_i(\mathbf{g}) \vdash \mathbf{B}_i(A)$ or $\mathbf{Ir}_i(\mathbf{g}) \vdash \mathbf{B}_i(\neg A)$; then we have the assertion by Axiom T. Let A be an atomic formula. Then, $g_1 \wedge g_2 \vdash A$ or $g_1 \wedge g_2 \vdash \neg A$. Then, $\mathbf{Ir}_i^o(\mathbf{g}) \vdash g_1 \wedge g_2$ by (6) and Axiom T. Thus, $\mathbf{Ir}_i^o(\mathbf{g}) \vdash A$ or $\mathbf{Ir}_i^o(\mathbf{g}) \vdash \neg A$.

Let A be nonatomic, and suppose the inductive hypothesis that decidability holds for the immediate subformulae of A. Let $A = C \supset D$. By the inductive hypothesis, decidability holds for C and D. Using this, we have $\mathbf{Ir}_i^o(\mathbf{g}) \vdash A$ or $\mathbf{Ir}_i^o(\mathbf{g}) \vdash \neg A$. Similar arguments apply to connectives $\wedge, \vee$ and $\neg$.

Let $A = \mathbf{B}_k(C)$. The hypothesis is: $\mathbf{Ir}_i^o(\mathbf{g}) \vdash C$ or $\mathbf{Ir}_i^o(\mathbf{g}) \vdash \neg C$. Let $\mathbf{Ir}_i^o(\mathbf{g}) \vdash C$. Then, $\mathbf{B}_k(\mathbf{Ir}_i^o(\mathbf{g})) \vdash \mathbf{B}_k(C)$. By IRA$_i^o$ and Axiom T, $\mathbf{Ir}_i^o(\mathbf{g}) \vdash \mathbf{B}_j(\mathbf{Ir}_i^o(\mathbf{g}))$ and $\mathbf{Ir}_i^o(\mathbf{g}) \vdash \mathbf{B}_i(\mathbf{Ir}_i^o(\mathbf{g}))$. Thus, $\mathbf{Ir}_i^o(\mathbf{g}) \vdash \mathbf{B}_k(C)$. Now, let $\mathbf{Ir}_i^o(\mathbf{g}) \vdash \neg C$. By the same arguments, we have $\mathbf{Ir}_i^o(\mathbf{g}) \vdash \mathbf{B}_k(\neg C)$, and, by Axiom D, $\mathbf{Ir}_i^o(\mathbf{g}) \vdash \neg\mathbf{B}_k(C)$.

Let $A = \mathbf{Ir}_k(C_1, C_2)$. The induction hypothesis is that decidability holds for C_1 and C_2. Now, suppose $\mathbf{Ir}_i^o(\mathbf{g}) \vdash C_1 \wedge C_2$. As remarked in the end of Section 2.2, $\mathbf{Ir}_i^o(\mathbf{g}) \vdash \mathbf{Ir}_j^o(\mathbf{g})$ and $\mathbf{Ir}_j^o(\mathbf{g}) \vdash \mathbf{Ir}_i^o(\mathbf{g})$. Hence, $\mathbf{Ir}_k^o(\mathbf{g}) \vdash C_k$ for $k = 1, 2$. Thus, $\mathbf{Ir}_k(\mathbf{g}) \vdash \mathbf{B}_k(C_k)$ for $k = 1, 2$. By Lemma 2.3 (1), $\mathbf{Ir}_k(\mathbf{g}) \vdash \mathbf{Ir}_k(C_1, C_2)$ for $k = 1, 2$. Since $\mathbf{Ir}_i^o(\mathbf{g}) \vdash \mathbf{Ir}_k(\mathbf{g})$ for $k = 1, 2$ by (6) and Axiom T, we have $\mathbf{Ir}_i^o(\mathbf{g}) \vdash \mathbf{Ir}_k(C_1, C_2)$.

Let $\mathbf{Ir}_i^o(\mathbf{g}) \vdash (\neg C_i) \wedge C_j$. By the same argument, we have $\mathbf{Ir}_i^o(\mathbf{g}) \vdash \mathbf{Ir}_i(\neg C_i; C_j)$. By Lemma 2.3.(5), $\mathbf{Ir}_i^o(\mathbf{g}) \vdash \neg\mathbf{Ir}_i(C_i; C_j)$.The same argument can be applied to the case of $\mathbf{Ir}_i^o(\mathbf{g}) \vdash C_i \wedge (\neg C_j)$ and $\mathbf{Ir}_i^o(\mathbf{g}) \vdash (\neg C_i) \wedge (\neg C_j)$.■

4 Prediction/Decision Making in the Logic IR^2

We give three axioms for player i's prediction/decision making, including some predictions about player j's decisions. We also assume the symmetric axioms for player i's prediction about player j's prediction/decision making. These lead to an infinite regress of those axioms, unless we stop at an arbitrary level. In this section, we show, for an interchangeable game, that the infinite regress of those axioms can be fully explicated, and obtain the decidability result.

4.1 Axioms for Prediction/Decision Making

We start with the following three axioms. These are described in the mind of player i, i.e., in the scope of $\mathbf{B}_i(\cdot)$;

$\mathrm{N0}_i$ **(Optimization against all predictions)**: $\wedge_{s\in S}[\mathrm{I}_i(s_i) \wedge \mathbf{B}_j(\mathrm{I}_j(s_j)) \supset \mathrm{bst}_i(s_i;s_j)]$.

$\mathrm{N1}_i$ **(Necessity of predictions)**: $\wedge_{s_i\in S_i}\langle \mathrm{I}_i(s_i) \supset \vee_{s_j\in S_j}\mathbf{B}_j(\mathrm{I}_j(s_j))\rangle$.

$\mathrm{N2}_i$ **(Predictability)**: $\wedge_{s_i\in S_i}\langle \mathrm{I}_i(s_i) \supset \mathbf{B}_j\mathbf{B}_i(\mathrm{I}_i(s_i))\rangle$.

For each $i = 1, 2$, let $\mathrm{N}_i = \mathrm{N0}_i \wedge \mathrm{N1}_i \wedge \mathrm{N2}_i$, and let $\mathbf{N} = (\mathrm{N}_1,\mathrm{N}_2)$.

The first axiom corresponds to Na_i. The second requires player i to have a prediction for his decision. It corresponds to the nonemptiness of E_1 and E_2 in Proposition 3.1, while $\mathrm{N1}_i$ allows both to be empty. The third states that in the mind of player i, his decision is correctly predicted by player j. We find a similar structure in Axiom IRA_i, but note that $\mathrm{N2}_i$ and IRA_i have different orders of applications of $\mathbf{B}_i$ and $\mathbf{B}_j$. Indeed, $\mathrm{I}_i(s_i)$ does not include the scope of $\mathbf{B}_i(\cdot)$, while $\mathbf{Ir}_i(\cdot,\cdot)$ can be regarded as including the outer $\mathbf{B}_i(\cdot)$, shown as in Lemma 2.2.

Axioms N_i and N_j are interdependent: N_i is assumed in the mind of player i, i.e., $\mathbf{B}_i(\mathrm{N}_i)$. Since $\mathbf{B}_i(\mathrm{N}_i)$ includes $\mathbf{B}_j(\mathrm{I}_j(s_j))$, player i needs to predict what j would choose. This prediction is made by the criterion $\mathbf{B}_i\mathbf{B}_j(\mathrm{N}_j)$. Then, $\mathbf{B}_i(\mathrm{I}_i(s_i))$ is included in $\mathbf{B}_i\mathbf{B}_j(\mathrm{N}_j)$, and it requires $\mathbf{B}_i\mathbf{B}_j\mathbf{B}_i(\mathrm{N}_i)$, and so on. These are captured by the infinite regress formula $\mathbf{Ir}_i(\mathbf{N}) = \mathbf{Ir}_i(\mathrm{N}_i;\mathrm{N}_j)$.

The above and their infinite regress $\mathbf{Ir}_i(\mathbf{N})$ in the logic IR^2 may be seen from Johansen's [10] interpretation of Nash equilibrium. This will be discussed in Section 6.

The infinite regress $\mathbf{Ir}_i(\mathrm{N}_i;\mathrm{N}_j)$ describes a necessary property for $\mathrm{I}_i(s_i)$ and $\mathrm{I}_j(s_j)$. We may find some candidates for such $\mathrm{I}_i(s_i)$'s $(i = 1, 2)$: for each $s_i \in S_i$,

$$A^*(s_i) := \vee_{t_j\in S_j}\mathbf{Ir}^o_i[\mathrm{bst}_i(s_i;t_j);\mathrm{bst}_j(t_j;s_i)]. \tag{12}$$

The nonepistemic content of $A^*(s_i)$ is given as $\varepsilon_0(A^*_i(s_i)) = \vee_{t_j\in S_j}\langle \mathrm{bst}_i(s_i;t_j)\ \wedge \mathrm{bst}_j(t_j;s_i)\rangle = \vee_{t_j\in S_j}\mathrm{nash}(s_i;t_j)$. That is, $\varepsilon_0(A^*_i(s_i))$ means "s_i is a Nash strategy". Also, in the logic $\mathrm{IR}^2(\mathrm{T})$ assuming Axiom T, we have $\vdash A^*(s_i) \equiv \vee_{t_j\in S_j}\mathbf{C}^*(\mathrm{nash}(s_i;t_j))$, i.e., $A^*(s_i)$ means "s_i is a common knowledge Nash strategy".

We have the following result, which will be proved in the end of this subsection.

Theorem 4.1. ***(Necessity)*** *For $i = 1, 2$,*

$$\mathbf{Ir}_i(\mathbf{N}) \vdash \mathbf{B}_i(I_i(s_i) \supset A^*_i(s_i)) \textit{ for all } s_i \in S_i. \tag{13}$$

That is, player i infers $A^*_i(s_i)$ as a necessary condition for a decision. By the theorem and Lemma 2.2, we have also $\mathbf{Ir}_i(\mathbf{N}) \vdash \mathbf{B}_i[\mathbf{B}_j(\mathrm{I}_j(s_j)) \supset \mathbf{B}_j(A^*_j(s_j))]$ for all $s_j \in S_j$; player i infers $\mathbf{B}_j(A^*_j(s_j)$ as a necessary conditions for a prediction. By Lemma 2.3.(1), we have, also, $\mathbf{Ir}_i(\mathbf{N}) \vdash \mathbf{Ir}_i[\mathrm{I}_i(s_i) \supset A^*_i(s_i);\mathrm{I}_j(s_j) \supset A^*_j(s_j)]$ for all $s \in S$. That is, those necessary conditions form an infinite regress, too. From now on, we talk about the statement of the form of (13).

With the remark on $\varepsilon_0(A^*_i(s_i))$, Theorem 4.1 may be interpreted as meaning that a Nash equilibrium is derived. However, our target is prediction/decision making by a player. A possible decision resulting from this process is described by $\mathrm{I}_i(s_i)$, and $A^*_i(s_i)$ is only a necessary condition for it. In addition, (13) is a purely solution-theoretic statement in the sense that it uses no specific structure of payoffs. Also, necessary condition (13) for $\mathrm{I}_i(s_i)$ does not give a positive answer to $\mathrm{I}_i(s_i)$ even if payoffs, e.g., $\mathbf{Ir}_i(g_1, g_2)$, are specified; that is, (13), or its contrapositive, may give only a negative decision $\neg\mathrm{I}_i(s_i)$ from $\neg A_i^*(s_i)$.

In Sections 4.2, 4.3, and Section 5.1, we discuss the converse of (13) under the assumption of $\mathbf{Ir}_i(g_1, g_2)$. Then we can discuss whether player i can make a decision or not.

We show the following lemma. Theorem 4.1 follows (2) of the lemma, and (1) does not need $\mathrm{N1}_i$. We write $\mathrm{N0}_i \wedge \mathrm{N2}_i$, $\mathrm{N0}_i \wedge \mathrm{N1}_i \wedge \mathrm{N2}_i$ as $\mathrm{N02}_i$, $\mathrm{N012}_i$ for $i = 1, 2$.

Lemma 4.1. *For $i = 1, 2$, and $s = (s_i; s_j) \in S$,*

(1): $\mathbf{Ir}_i^o[N02_i;N02_j] \vdash I_i(s_i) \wedge \mathbf{B}_j(I_j(s_j)) \supset \mathbf{Ir}_i^o[bst_i(s_i; s_j);bst_j(s_i; s_j)]$;

(2): $\mathbf{Ir}_i^o[N012_i;N012_j] \vdash I_i(s_i) \supset A^*(s_i)$.

Proof. (1): Let $\theta_i(s_i, s_j) := \mathbf{Ir}_i^o[\mathrm{N02}_i,\mathrm{N02}_j] \wedge \mathrm{I}_i(s_i) \wedge \mathbf{B}_j(\mathrm{I}_j(s_j))$. Here, we show, for $i = 1, 2$,

$$\vdash \theta_i(s_i, s_j) \supset \mathrm{bst}_i(s_i; s_j) \wedge \mathbf{B}_j(\mathrm{bst}_j(s_j; s_i)) \wedge \mathbf{B}_j\mathbf{B}_i(\theta_i(s_i, s_j)). \tag{14}$$

Once this is shown, we have, by Lemma 2.4.(2), $\vdash \theta_i(s_i, s_j) \supset \mathbf{Ir}_i^o[\mathrm{bst}_i(s_i; s_j),\mathrm{bst}_j(s_i; s_j)]$, which implies the assertion.

The first part, $\vdash \theta_i(s_i, s_j) \supset \mathrm{bst}_i(s_i; s_j)$, of (14) comes from $\mathrm{N0}_i$ and $\mathrm{I}_i(s_i) \wedge \mathbf{B}_j(\mathrm{I}_j(s_j))$. Consider the second part. Since $\vdash \theta_i(s_i, s_j) \supset \mathbf{B}_j(\mathrm{N02}_j)$ and $\vdash \mathbf{B}_j(\mathrm{N02}_j) \wedge \mathbf{B}_j(\mathrm{I}_j(s_j)) \wedge \mathbf{B}_j\mathbf{B}_i(\mathrm{I}_i(s_i)) \supset \mathbf{B}_j(\mathrm{bst}_j(s_j; s_i))$, we have $\vdash \theta_i(s_i, s_j) \wedge \mathbf{B}_j(\mathrm{I}_j(s_j)) \wedge \mathbf{B}_j\mathbf{B}_i(\mathrm{I}_i(s_i)) \supset \mathbf{B}_j(\mathrm{bst}_j(s_j; s_i))$. The $\mathbf{B}_j(\mathrm{I}_j(s_j))$ is included in $\theta_i(s_i, s_j)$, and the $\mathbf{B}_j\mathbf{B}_i(\mathrm{I}_i(s_i))$ is derived from $\mathrm{I}_i(s_i)$ in $\theta_i(s_i, s_j)$ by $\mathrm{N2}_i$. Hence, $\vdash \theta_i(s_i, s_j) \supset \mathbf{B}_j(\mathrm{bst}_j(s_j; s_i))$. Now, consider the third part of (14). By Lemma 2.4.(1), $\vdash \mathbf{Ir}_i^o[\mathrm{N02}_i;\mathrm{N02}_j] \supset \mathbf{B}_j\mathbf{B}_i(\mathbf{Ir}_i^o[\mathrm{N02}_i;\mathrm{N02}_j])$. Using $\mathrm{N2}_i$, we have $\vdash \mathbf{Ir}_i^o[\mathrm{N02}_i;\mathrm{N012}_j] \wedge \mathrm{I}_i(s_i) \supset \mathbf{B}_j\mathbf{B}_i(\mathrm{I}_i(s_i))$, and, using $\mathbf{B}_j(\mathrm{N2}_j)$ in $\mathbf{Ir}_i^o[\mathrm{N02}_i;\mathrm{N02}_j]$, we have $\vdash \mathbf{Ir}_i^o[\mathrm{N02}_i;\mathrm{N02}_j] \wedge \mathbf{B}_j(\mathrm{I}_j(s_j)) \supset \mathbf{B}_j\mathbf{B}_i\mathbf{B}_j(\mathrm{I}_j(s_j))$. Hence, we have $\vdash \theta_i(s_i, s_j) \supset \mathbf{B}_j\mathbf{B}_i(\theta_i(s_i, s_j))$.

(2): It follows from (1) that $\mathbf{Ir}_i^o[\mathrm{N02}_i;\mathrm{N02}_j] \vdash \mathrm{I}_i(s_i) \wedge \mathbf{B}_j(\mathrm{I}_j(s_j)) \supset \vee_{t_j \in S_j} \mathbf{Ir}_i^o[\mathrm{bst}_i(s_i; t_j);\mathrm{bst}_j(t_j; s_i)]$. This is equivalent to $\mathbf{Ir}_i^o[\mathrm{N02}_i;\mathrm{N02}_j] \vdash \mathbf{B}_j(\mathrm{I}_j(s_j)) \supset (\mathrm{I}_i(s_i) \supset A_i^*(s_i))$. Hence $\mathbf{Ir}_i^o[\mathrm{N02}_i;\mathrm{N02}_j] \vdash \vee_{t_j \in S_j} \mathbf{B}_j(\mathrm{I}_j(t_j)) \supset (\mathrm{I}_i(s_i) \supset A_i^*(s_i))$. Adding $\mathrm{N1}_i$ to $\mathbf{Ir}_i^o[\mathrm{N02}_i,\mathrm{N02}_j]$, we delete the first disjunctive formula, i.e., $\mathbf{Ir}_i^o[\mathrm{N012}_i;\mathrm{N012}_j] \vdash \mathrm{I}_i(s_i) \supset A_i^*(s_i)$.■

4.2 Choice of the deductive weakest formulae for $\mathbf{N}_i$ and $\mathbf{N}_j$

There are some concrete formulae $A_i(s_i)$ and $A_j(s_j)$ enjoying the properties described by N_i and N_j. We, however, find some unintended candidates for those axioms. For example, the contradictory formulae $\bot(s_i) := \neg(p_i(s_i) \supset p_i(s_i))$, $s_i \in S_i$ are trivial candidates for them, where $p_i(s_i) := \vee_{t_j \in S_j} \mathrm{Pr}_i(s_i, t_j; s_i, t_j)$. Indeed, the class of formulae $\{\bot(s_i)\}_{s_i \in S_i}$ makes $\mathrm{N012}_i$ trivially hold with the substitution of $\bot(s_i)$ for each $\mathrm{I}_i(s_i)$ in N_i. We need to choose a class of formulae $\mathcal{A}_i = \{A_i(s_i)\}_{s_i \in S_i}$ and $\mathcal{A}_j = \{A_i(s_j)\}_{s_j \in S_j}$ having *only* the properties N_i and N_j.

Let $\mathcal{A} = (\mathcal{A}_i; \mathcal{A}_j)$ be a pair of *candidate families* indexed by $s_i \in S_i$ and $s_j \in S_j$. Let $\mathrm{N}_i(\mathcal{A})$ be the formula obtained from N_i by substituting $(A_1(s_1), A_2(s_2))$ for $(\mathrm{I}_1(s_1),\mathrm{I}_2(s_2))$ for each $s = (s_1, s_2) \in S$. We denote the following formula by $\mathrm{WF}_i(\mathcal{A})$:

$$\begin{aligned} \mathrm{N}_i(\mathcal{A}) \wedge \mathbf{B}_j(\mathrm{N}_j(\mathcal{A})) \wedge [\wedge_{s \in S}\{\mathrm{I}_i(s_i) \wedge \mathbf{B}_j(\mathrm{I}_j(s_j)) \ &\supset \ A_i(s_i) \wedge \mathbf{B}_j(A_j(s_j))\}] \\ &\supset \ \wedge_{s_i \in S_i}\{A_i(s_i) \supset \mathrm{I}_i(s_i)\}. \end{aligned} \tag{15}$$

Let $\mathbf{WF}(\mathcal{A}) = (\mathrm{WF}_1(\mathcal{A}),\mathrm{WF}_2(\mathcal{A}))$. We denote, by $\mathbf{Ir}_i(\mathbf{WF})$, the set $\{\mathbf{Ir}_i(\mathbf{WF}(\mathcal{A})) : \mathcal{A} = (\mathcal{A}_1, \mathcal{A}_2)$ is a pair of candidate families of formulae$\}$.

The formula $\mathrm{WF}_i(\mathcal{A})$ contains an additional premise $\wedge_{s\in S}\{\mathrm{I}_i(s_i) \wedge \mathbf{B}_j(\mathrm{I}_j(s_j)) \supset A_i(s_i) \wedge \mathbf{B}_j(A_j(s_j))\}$. A sole use of $\mathrm{WF}_i(\mathcal{A})$ is not meaningful since $\mathrm{I}_i(s_i)\wedge\mathbf{B}_j(\mathrm{I}_j(s_j))$ have no properties, yet. We assume $\mathbf{Ir}_i(\mathbf{WF})$ together with $\mathbf{Ir}_i(\mathbf{N})$. Then, the premise avoids some double cross of $\mathrm{I}_i(s_i) \wedge \mathbf{B}_j(\mathrm{I}_j(s_j)$ and $A_i(s_i) \wedge \mathbf{B}_j(A_j(s_j))$, both of which satisfy N_i and N_j.

The additional premise is of the same nature as the term "maximal" used in the definition of a subsolution in Section 3; we cannot use the term "largest" for a subsolution. If we drop the additional premise, (15) becomes

$$WF_i^+(\mathcal{A}) = \mathrm{N}_i(\mathcal{A}) \wedge \mathbf{B}_j(\mathrm{N}_j(\mathcal{A})) \supset \wedge_{s_i\in S_i}\{A_i(s_i) \supset \mathrm{I}_i(s_i)\}. \tag{16}$$

This is stronger than $WF_i(\mathcal{A})$, since it has a weaker premise. This strengthening $WF_i^+(\mathcal{A})$ works only for an interchangeable game, but not for an uninterchangeable game, while $WF_i(\mathcal{A})$ in (15) works for any game.

We study implications from $\{\mathbf{Ir}_i(\mathbf{N})\} \cup \mathbf{Ir}_i(\mathbf{WF})$ under the infinite regress of formalized payoffs $\mathbf{Ir}_i(\mathbf{g}) = \mathbf{Ir}_i(g_i; g_j)$. The entire set of axioms is denoted by $\Delta_i := \{\mathbf{Ir}_i(\mathbf{g}), \mathbf{Ir}_i(\mathbf{N})\} \cup \mathbf{Ir}_i(\mathbf{WF})$. We have the following lemma, which will be proved in the proof of Lemma 5.1.

Lemma 4.2. ***(Consistency of the belief set)*** *Δ_i is consistent for any game G.*

In fact, $\Delta_i^+ = \{\mathbf{Ir}_i(\mathbf{g}), \mathbf{Ir}_i(\mathbf{N})\}\cup\mathbf{Ir}_i(\mathbf{WF}^+)$ is consistent if and only if G is an interchangeable game, and Δ_i^+ is equivalent to Δ_i for any interchangeable G.

4.3 Characterization and decidability for interchangeable games

Here, we show that our axioms characterize the possible final decisions for an interchangeable game. A proof of this theorem is given in the end of this subsection.

Theorem 4.2. ***(Characterization I)*** *Let G be an interchangeable game and $\mathbf{g} = (g_1, g_2)$ its formalized payoffs. Then, for $i = 1, 2$,*

$$\Delta_i \vdash \mathbf{B}_i(I_i(s_i) \equiv A_i^*(s_i)) \text{ for all } s_i \in S_i. \tag{17}$$

This is interpreted as meaning that player i infers from his beliefs Δ_i that his possible decision and prediction are fully expressed by $A_i^*(s_i)$ and $A_j^*(s_j)$ for an interchangeable game G. As remarked above, in the logic $\mathrm{IR}^2(\mathrm{T})$, $A_i^*(s_i)$ is equivalent to $\vee_{t_j\in S_j}\mathbf{C}^*(\mathrm{Nash}(s_i; t_j))$, and Theorem 4.2 becomes $\Delta_i \vdash \mathrm{I}_i(s_i) \equiv \vee_{t_j\in S_j}\mathbf{C}^*(\mathrm{Nash}(s_i; t_j))$. That is, a possible decision s_i is the Nash strategy with common knowledge. This corresponds to the result given in Kaneko [12].

Then, player i can even decide whether a given strategy s_i is a final decision for him or not. This is described by the following theorem.

Theorem 4.3. ***(Positive or negative decisions)*** *Let G be an interchangeable game and $\mathbf{g} = (g_1, g_2)$ its formalized payoffs. Then, for $i = 1, 2$ and each $s_i \in S_i$,*

$$\text{either } \Delta_i \vdash \mathbf{B}_i(I_i(s_i)) \text{ or } \Delta_i \vdash \mathbf{B}_i(\neg I_i(s_i)). \tag{18}$$

Proof. We show (18). Since $\mathrm{bst}_i(s_i; s_j)$ is a nonepistemic game formula for i, it follows from Lemma 3.1.(2) that $\mathbf{Ir}_i^o(\mathbf{g}) \vdash \mathbf{Ir}_i^o[\mathrm{bst}_i(s_i; s_j); \mathrm{bst}_j(s_j; s_i)]$ if and only if $g_i \vdash \mathrm{bst}_i(s_i; s_j)$ and $g_j \vdash \mathrm{bst}_j(s_j; s_i)$. By Lemma 3.1.(1), $\mathbf{Ir}_i^o(\mathbf{g}) \vdash \mathbf{Ir}_i^o[\mathrm{bst}_i(s_i; s_j);\mathrm{bst}_j(s_j; s_i)]$ or $\mathbf{Ir}_i^o(\mathbf{g}) \vdash \neg\mathbf{Ir}_i^o[\mathrm{bst}_i(s_i; s_j); \mathrm{bst}_j(s_j; s_i)]$.

If s_i is a Nash strategy for G, then $\mathbf{Ir}_i^o(\mathbf{g}) \vdash \vee_{t_j}\mathbf{Ir}_i^o[\mathrm{bst}_i(s_i; t_j);\mathrm{bst}_j(t_j; s_i)]$, i.e., $\mathbf{Ir}_i^o(\mathbf{g}) \vdash A_i^*(s_i)$, and otherwise, Then, $\mathbf{Ir}_i^o(\mathbf{g}) \vdash \neg \vee_{t_j} \mathbf{Ir}_i^o[\mathrm{bst}_i(s_i; t_j);\mathrm{bst}_j(t_j; s_i)]$, i.e., $\mathbf{Ir}_i^o(\mathbf{g}) \vdash \neg A_i^*(s_i)$. Thus, we have $\mathbf{Ir}_i(\mathbf{g}) \vdash \mathbf{B}_i(A_i^*(s_i))$ or $\mathbf{Ir}_i(\mathbf{g}) \vdash \mathbf{B}_i(\neg A_i^*(s_i))$. By (17), we have $\Delta_i \vdash \mathbf{B}_i(\mathrm{I}_i(s_i))$ or $\Delta_i \vdash \mathbf{B}_i(\neg\mathrm{I}_i(s_i))$.■

By Theorem 4.3 and Lemma 2.3.(1), we have also, for each strategy $s_j \in S_j$,

$$\text{either } \Delta_i \vdash \mathbf{B}_i\mathbf{B}_j(\mathrm{I}_j(s_j)) \text{ or } \Delta_i \vdash \mathbf{B}_i\mathbf{B}_j(\neg\mathrm{I}_j(s_j)). \tag{19}$$

Thus, player i can predict whether a given strategy for j is a decision for him for not.

Since $\varepsilon_0 A_i^*(s_i) = \vee_{t_j \in S_j}\mathrm{nash}(s_i; t_j)$, the positive or negative decision in (18) corresponds to whether s_i is a Nash strategy or not. For the negative case, we need to add only $\mathbf{Ir}_i(\mathbf{g})$ to $\mathbf{Ir}_i(\mathbf{N})$ of Theorem 4.1, that is, if s_i is not a Nash strategy, then

$$\mathbf{Ir}_i(\mathbf{g}), \mathbf{Ir}_i(\mathbf{N}) \vdash \mathbf{B}_i(\neg\mathrm{I}_i(s_i)). \tag{20}$$

This result is independent of the interchangeability of the game G. For the positive case, we need the full set $\Delta_i = \{\mathbf{Ir}_i(\mathbf{g}), \mathbf{Ir}_i(\mathbf{N})\} \cup \mathbf{Ir}_i(\mathbf{WF})$ and the interchangeability of G.

Since Table 1.1 is an interchangeable game, Theorem 4.3 is applied to it, and the belief set Δ_1 recommends strategy $\mathbf{s}_{12}$ as a positive decision, but $\mathbf{s}_{11}$, $\mathbf{s}_{13}$ as negative decisions. By (19), player 2 would choose $\mathbf{s}_{21}$, and denies the others. Table 1.2 has an uninterchangeable game; Theorem 4.2 is not applicable. (20) is applied to Table 1.3, and recommends any strategy as a negative decision.

Let us prove Theorem 4.2. First, we show the following lemma.

Lemma 4.3. *Let G be a 2-person game.*

(0): Let G be interchangeable. Then, $\mathbf{Ir}_i^o(g_i; g_j) \vdash A_i^*(s_i) \wedge \mathbf{B}_j(A_j^*(s_j)) \supset bst_i(s_i; s_j)$.

(1): $\vdash A_i^*(s_i) \supset \vee_{t_j \in S_j}\mathbf{B}_j(A_j^*(t_j))$.

(2): $\vdash A_i^*(s_i) \supset \mathbf{B}_j\mathbf{B}_i(A_i^*(s_i))$.

Proof. (0): Since $\mathrm{bst}_i(s_i; s_j)$ is a game formula for $i = 1, 2$, we have, for each $s \in S$, $\mathbf{Ir}_i^o(g_i; g_j) \vdash \mathbf{Ir}_i^o(\mathrm{bst}_i(s_i; s_j);\mathrm{bst}_j(s_j; s_i))$ or $\mathbf{Ir}_i^o(g_i; g_j) \vdash \neg\mathbf{Ir}_i^o(\mathrm{bst}_i(s_i; s_j); \mathrm{bst}_j(s_j; s_i))$ by Theorem 3.1. Hence, for each $s_i \in S_i$, $\mathbf{Ir}_i^o(g_i; g_j) \vdash A_i^*(s_i)$ or $\mathbf{Ir}_i^o(g_i; g_j) \vdash \neg A_i^*(s_i)$. Using Lemma 2.2, we have, for each $s_j \in S_j$, $\mathbf{Ir}_i^o(g_i; g_j) \vdash \mathbf{B}_j(A_j^*(s_j))$ or $\mathbf{Ir}_i^o(g_i; g_j) \vdash \neg\mathbf{B}_j(A_j^*(s_j))$. Also, for each $s \in S$, $\mathbf{Ir}_i^o(g_i; g_j) \vdash \mathrm{bst}_i(s_i; s_j)$ or $\mathbf{Ir}_i^o(g_i; g_j) \vdash \neg\mathrm{bst}_i(s_i; s_j)$. Thus, $\mathbf{Ir}_i^o(g_i; g_j) \vdash A_i^*(s_i) \wedge \mathbf{B}_j(A_j^*(s_j)) \supset \mathrm{bst}_i(s_i; s_j)$ or $\mathbf{Ir}_i^o(g_i; g_j) \vdash \neg[A_i^*(s_i) \wedge \mathbf{B}_j(A_j^*(s_j)) \supset \mathrm{bst}_i(s_i; s_j)]$. If the latter held, then, applying the epistemic eraser ε_0 to this, we would have $g_i \wedge g_j \vdash \neg[(\vee_{t_j \in S_j}\mathrm{nash}(s_i, t_j)) \wedge (\vee_{t_i \in S_i}\mathrm{nash}(s_j, t_i)) \supset \mathrm{bst}_i(s_i; s_j)]$, which is impossible since G is an interchangeable game. Hence, we have the assertion.

(1): By Lemma 2.2, we have $\vdash \mathbf{Ir}_i^o[\mathrm{bst}_i(s_i; s_j);\mathrm{bst}_j(s_j; s_i)] \supset \mathbf{B}_j(\mathbf{Ir}_j^o[\mathrm{bst}_j(s_j; s_i); \mathrm{bst}_i(s_i; s_j)])$. Hence, $\vdash \mathbf{Ir}_i^o[\mathrm{bst}_i(s_i; s_j); \mathrm{bst}_j(s_j; s_i)] \supset \mathbf{B}_j(\vee_{t_i \in S_i}\mathbf{Ir}_j^o[\mathrm{bst}_j(s_j; t_i); \mathrm{bst}_i(s_i; t_j)])$, i.e., $\vdash \mathbf{Ir}_i^o[\mathrm{bst}_i(s_i; s_j); \mathrm{bst}_j(s_j; s_i)] \supset \mathbf{B}_j(A_j^*(s_j))$. Hence, $\vdash \mathbf{Ir}_i^o[\mathrm{bst}_i(s_i; s_j);\mathrm{bst}_j(s_j; s_i)] \supset \vee_{t_j \in S_j}\mathbf{B}_j(A_j^*(t_j))$. Then, $\vdash \vee_{t_j \in S_j}\mathbf{Ir}_i^o[\mathrm{bst}_i(s_i; t_j),\mathrm{bst}_j(t_j; s_i)] \supset \vee_{t_j \in S_j}\mathbf{B}_j(A_j^*(t_j))$, i.e., $\vdash A_i^*(s_i) \supset \vee_{t_j \in S_j}\mathbf{B}_j(A_j^*(t_j))$.

(2): Since $\vdash \mathbf{Ir}_i^o[\mathrm{bst}_i(s_i;s_j);\mathrm{bst}_j(s_j;s_i)] \supset \mathbf{B}_j(\mathbf{Ir}_j^o[\mathrm{bst}_j(s_j;s_i);\mathrm{bst}_i(s_i;s_j)])$ and $\vdash \mathbf{B}_j(\mathbf{Ir}_j^o[\mathrm{bst}_j(s_j;s_i);\mathrm{bst}_i(s_i;s_j)]) \supset \mathbf{B}_j\mathbf{B}_i(\mathbf{Ir}_i^o[\mathrm{bst}_i(s_i;s_j);\ \mathrm{bst}_j(s_j;s_i)])$, we have $\vdash \mathbf{Ir}_i^o[\mathrm{bst}_i(s_i;s_j);\mathrm{bst}_j(s_j;s_i)] \supset \mathbf{B}_j\mathbf{B}_i(\mathbf{Ir}_i^o[\mathrm{bst}_i(s_i;s_j);\mathrm{bst}_j(s_j;s_i)])$. We take disjunctions from the latter to the former with respect to s_j, and have $\vdash \vee_{t_j\in S_j}\mathbf{Ir}_i^o[\mathrm{bst}_i(s_i;t_j),\mathrm{bst}_j(t_j;s_i)] \supset \vee_{t_j\in S_j}\mathbf{B}_j\mathbf{B}_i(\mathbf{Ir}_i^o[\mathrm{bst}_i(s_i;t_j),\mathrm{bst}_j(t_j;s_i)])$. Then, the former is $A_i^*(s_i)$, and the latter implies $\mathbf{B}_j\mathbf{B}_i(\vee_{t_j\in S_j}\mathbf{Ir}_i^o[\mathrm{bst}_i(s_i;t_j),\mathrm{bst}_j(t_j;s_i)])$, i.e., $\mathbf{B}_j\mathbf{B}_i(A_i^*(s_i))$.$\blacksquare$

Proof of Theorem 4.2. It follows from Lemma 4.3 that $\mathbf{Ir}_i^o(g_i;g_j) \vdash \mathrm{N}_i(\mathcal{A}^*)$ for $i=1,2$. Hence, $\mathbf{Ir}_i^o(g_i;g_j) \vdash \mathrm{N}_i(\mathcal{A}^*) \wedge \mathbf{B}_j(\mathrm{N}_j(\mathcal{A}^*))$. It follows from Theorem 4.1 that $\mathbf{Ir}_i^o(\mathrm{N}_i;\mathrm{N}_j) \vdash \wedge_{s\in S}[\mathrm{I}_i(s_i) \wedge \mathbf{B}_j(\mathrm{I}_j(s_j)) \supset A_i^*(s_i) \wedge \mathbf{B}_j(A_j^*(s_j))]$. We have $\mathbf{Ir}_i^o(g_i;g_j), \mathbf{Ir}_i^o(\mathbf{N}), \mathbf{Ir}_i^o(\mathbf{WF}) \vdash [A_i^*(s_i) \supset \mathrm{I}_i(s_i)] \wedge [\mathbf{B}_j(A_i^*(s_i)) \supset \mathbf{B}_j(\mathrm{I}_j(s_j))]$. Hence, $\mathbf{Ir}_i^o(g_i;g_j), \mathbf{Ir}_i^o(\mathbf{N}), \mathbf{Ir}_i^o(\mathbf{WF}) \vdash \mathbf{Ir}_i^o[A_i^*(s_i) \supset \mathrm{I}_i(s_i); A_i^*(s_j) \supset \mathrm{I}_i(s_j))]$. Using Theorem 4.1 and Lemma 2.3.(3), we have $\mathbf{Ir}_i(g_i;g_j), \mathbf{Ir}_i(\mathbf{N}), \mathbf{Ir}_i(\mathbf{WF}) \vdash \mathbf{Ir}_i[A_i^*(s_i) \equiv \mathrm{I}_i(s_i); A_i^*(s_j) \equiv \mathrm{I}_i(s_j))]$.$\blacksquare$

5 Undecidability for Uninterchangeable Games

The situation for an uninterchangeable game differs entirely from that for an interchangeable game. For an interchangeable G, we show the undecidability result that for some strategy s_i for player i, he cannot infer from his belief set $\mathbf{\Delta}_i = \{\mathbf{Ir}_i(g_i;g_j), \mathbf{Ir}_i(\mathbf{N})\} \cup \mathbf{Ir}_i(\mathbf{WF})$ whether s_i is a final decision or not. We give three other results related to this theorem.

5.1 Undecidability theorem and related theorems

Theorem 5.1. ***(Undecidability of prediction/decision making)*** *Let G be an uninterchangeable game, $\mathbf{g} = (g_1, g_2)$ its formalized payoffs, and $i = 1,2$. Then, there is an $s_i \in S_i$ such that*

$$\textit{neither } \mathbf{\Delta}_i \vdash \mathbf{B}_i(I_i(s_i)) \ \textit{ nor } \ \mathbf{\Delta}_i \vdash \mathbf{B}_i(\neg I_i(s_i)). \tag{21}$$

This will be proved in Section 5.2. First, we note that we have some s_j so that neither $\mathbf{\Delta}_i \vdash \mathbf{B}_i\mathbf{B}_j(\mathrm{I}_j(s_j))$ nor $\mathbf{\Delta}_i \vdash \mathbf{B}_i\mathbf{B}_j(\neg\mathrm{I}_j(s_j))$, i.e., player i cannot predict whether s_j is a decision or not for player j. This is also obtained in the proof of Theorem 5.1. Now, we concentrate on (21) for player i.

The following theorem states that the decision $\mathrm{I}_i(s_i)$ cannot be expressed in terms of a concrete formula if (21) holds for s_i, which is proved in Section 5.2.

Theorem 5.2. ***(No-formula)*** *Let G be an uninterchangeable game, $\mathbf{g} = (g_1, g_2)$ its formalized payoffs, and $i = 1,2$. Let $s_i \in S_i$ be a strategy for which (21) holds. Then, there is no game formula A_i such that*

$$\mathbf{\Delta}_i \vdash \mathbf{B}_i(I_i(s_i) \equiv A_i). \tag{22}$$

Theorems 5.1 and 5.2 hold even for IR^2 with the additional axioms T, 4 and 5. We will prove Theorem 5.2 for $\mathrm{IR}^2(\mathrm{T})$, which implies the result for IR^2.

The undecidability result differs from the negative result for a game with no Nash equilibria: For such a game, Theorem 4.3 states that player i can deny any strategy for his decision. In this case, he may think about some other criterion such as the default criterion that the first strategy for him should be chosen. However, undecidability means that he can not reach such a conclusion.

The negative decision given in (20) holds for a non-Nash strategy s_i for any game G. Hence, s_i for (21) is a Nash strategy. In fact, it is a sufficient (in fact, necessary, too) condition for (21) that

$$s_i \text{ is a Nash strategy but } s_i \notin F_i \text{ for some subsolution } F_1 \times F_2, \tag{23}$$

which is shown in Lemma 5.1. The battle of the sexes (Table 1.2) has two subsolutions $\{(\mathbf{s}_{11}, \mathbf{s}_{21})\}$, $\{(\mathbf{s}_{12}, \mathbf{s}_{22})\}$. Since (23) holds for each of $\mathbf{s}_{i1}$ and $\mathbf{s}_{i2}$, we have undecidability (21) for both strategies of both players.

Table 5.1

	$\mathbf{s}_{21}$	$\mathbf{s}_{22}$
$\mathbf{s}_{11}$	${}^{F^1}(1,1)^{F^2}$	$(0,1)^{F^2}$
$\mathbf{s}_{12}$	${}^{F^1}(1,0)$	$(0,0)$

Even when G is uninterchangeable, there may be some case where player i has a positive decision. Table 5.1 has two subsolutions $F^1 = \{(\mathbf{s}_{11}, \mathbf{s}_{21}), (\mathbf{s}_{12}, \mathbf{s}_{21})\}$ and $F^2 = \{(\mathbf{s}_{11}, \mathbf{s}_{21}), (\mathbf{s}_{11}, \mathbf{s}_{22})\}$. Since $(\mathbf{s}_{11}, \mathbf{s}_{21})$ belongs to both subsolutions, (23) does not hold for $\mathbf{s}_{i1}$.

To consider what would happen when the subsolutions have a nonempty intersection, we extend Theorem 4.2 to an uninterchangeable game G. Let G be any game with its subsolutions $F^1, ..., F^k$. We denote the intersection $\cap_{l=1}^k F^l$ by $\hat{F}$. We stipulate that $k = 0$ and $\hat{F} = \emptyset$ if G has no Nash equilibria. If G is solvable, then $k = 1$ and F^1 is the set of all Nash equilibria $E(G)$. We note that this intersection $\hat{F}$ satisfies interchangeability; so it can be written as $\hat{F}_1 \times \hat{F}_2$.

Here, we modify the target formulae $\{A_i^*(s_i)\}_{i \in S_i}, i = 1, 2$,as follows:

$$A^{**}(s_i) := \vee_{t_j \in \hat{F}_j} \mathbf{Ir}_i^o[\mathrm{bst}_i(s_i; t_j); \mathrm{bst}_j(t_j; s_i)]. \tag{24}$$

This differs from $A^*(s_i)$ with the domain of disjunction $\hat{F}_j$ instead of S_j. In this sense, it depends upon the specification of the payoff functions.

We define the candidate formulae $\mathcal{C}_i = \{C_i^*(s_i)\}_{s_i \in S_i}, i = 1, 2$ as follows:

$$C_i^*(s_i) = \begin{cases} A_i^{**}(s_i) & \text{if } s_i \in \hat{F}_i \\ A_i^*(s_i) & \text{if } s_i \notin E(G)_i \\ \mathrm{I}_i(s_i) & \text{otherwise.} \end{cases} \tag{25}$$

Then, we have the following characterization theorem, which will be proved in Section 5.2.

Theorem 5.3. ***(Characterization II)*** *Let G be any game with its subsolutions $F^1, ..., F^k$, $\mathbf{g} = (g_1, g_2)$ its formalized payoffs, and $i = 1, 2$. Then, $\vdash \mathbf{B}_i(I_i(s_i) \equiv C_i^*(s_i))$ for all $s_i \in S_i$.*

The following theorem is a corollary.

Theorem 5.4. ***(Positive Decision)*** *Let G be any game, $\mathbf{g} = (g_1, g_2)$ its formalized payoffs, and $i = 1, 2$. Then, for all $s_i \in S_i$, $\Delta_i \vdash \mathbf{B}_i(I_i(s_i))$ if and only if $s_i \in \hat{F}_i$.*

This has various implications: When G has no Nash equilibria, i.e., $\hat{F} = \emptyset$, Δ_i gives no positive decisions; When G is solvable, it gives a positive decision. When G has a multiple subsolutions, there are two cases; if $\hat{F} = \emptyset$, then it gives no positive decision; and if $\hat{F} \neq \emptyset$, it gives a positive decision, i.e., $s_i \in \hat{F}_i$.

It may be informative to state the semantic counterpart of Theorem 5.1, but since this goes to a sidetrack, we do not give a proof.

Theorem 5.5. ***(Semantic counterpart):*** *Let G be a game with $E(G) \neq \emptyset$ and $\mathbf{g} = (g_1, g_2)$ its associated formalized payoffs. Let $M = (\langle W; R_1, R_2 \rangle, \tau)$ be any KD-model and w any world in W. Suppose that $(M, w) \models \mathbf{Ir}_i(\mathbf{g}) \wedge \mathbf{Ir}_i(\mathbf{N})$ and $(M, w) \models \mathbf{Ir}_i(\mathbf{WF}(\mathcal{A}))$ for all $\mathcal{A}$. Then, there exists a subsolution $F = F_1 \times F_2$ in G such that*

$$(M, w) \models \mathbf{Ir}_i(I_1(s_1), I_2(s_1)) \text{ for any } s \in F, \tag{26}$$

$$(M, w) \models \mathbf{Ir}_i(\neg I_1(s_1), \neg I_2(s_1)) \text{ for any } s \in (S_1 - F_1) \times (S_2 - F_2)\ . \tag{27}$$

When G has no Nash equilibria, the theorem is modified as stating the the claimed F is empty: Only (27) is applied.

Theorem 5.5 describes how a subsolution F occurs in one model. It states that in the world $w \in W$, only s_i and s_j from the subsolution F are a decision and a prediction for player i. From the viewpoint of a single model, this resolves the difficulty caused by our undecidability result. However, we take the viewpoint that player i's inference is described in the formal system of IR^2. The soundness/completeness theorem (Theorem 2.1) implies that for each model, a subsolution may be a possible solution but the choice of a model remains.

5.2 Proof of the theorems

We stipulate that when $E(G) = \emptyset$, then the subsolution F is empty and $F_1 = F_2 = \emptyset$. Lemma 4.2 follows this lemma and soundness for IR^2.

Lemma 5.1. *Let G be any game. Then, for any subsolution $F = F_1 \times F_2$ in G, there is a KD-model $M = (\langle W; R_1, R_2 \rangle, \tau)$ and a world $w \in W$ such that*

$$(M, w) \models \mathbf{Ir}_i(\mathbf{g}) \wedge \mathbf{Ir}_i(\mathbf{N}) \text{ and } (M, w) \models \mathbf{Ir}_i(\mathbf{WF}(\mathcal{A})) \text{ for all } \mathcal{A}; \tag{28}$$

$$\text{for any } s_i \in S_i,\ (M, w) \models \mathbf{B}_i(I_i(s_i)) \Leftrightarrow (M, w) \models I_i(s_i) \Leftrightarrow s_i \in F_i. \tag{29}$$

Proof. We construct a model $M = (\langle W; R_1, R_2 \rangle, \tau)$ satisfying (28) and (29). Let $F = F_1 \times F_2$ be a subsolution. Let $\langle W; R_1, R_2 \rangle$ be the frame given by $W = \{w\}$ and $R_k = \{(w, w)\}$ for $k = 1, 2$, i.e., it has a single world, and R_k is reflexive. Hence, this is a frame for Axiom T (and 4, 5), too. Define τ by, for $k = 1, 2$,

$$\text{for any } s; s' \in S,\ \tau(\mathrm{PR}_k(s; s')) = \top \Leftrightarrow h_k(s) \geq h_k(s'); \tag{30}$$

$$\tau(w, \mathrm{I}_k(s_k)) = \top \Leftrightarrow s_k \in F_k. \tag{31}$$

That is, the preferences true relative to h_k are given by τ; and $\mathrm{I}_k(s_k)$ is true if and only if $s_k \in F_k$. By (30), we have $(M, w) \models g_1 \wedge g_2$. Also, since $W = \{w\}$, we have, for any formula C and $k = 1, 2$,

$$(M, w) \models C \ \Leftrightarrow (M, w) \models \mathbf{B}_k(C). \tag{32}$$

Now, because F is a subsolution and $(M,w) \models g_1 \wedge g_2$, it follows that $(M,w) \models \mathrm{bst}_i(s_i;s_j)$ for all $(s_i;s_j) \in F$ and for $i=1,2$. Thus, $(M,w) \models \mathrm{N0}_i$. Also, $(M,w) \models \mathrm{N1}_i$ by (31), and $(M,w) \models \mathrm{N2}_i$ by $W=\{w\}$. Thus, $(M,w) \models \mathbf{Ir}_i(\mathbf{N})$ for both $i=1,2$.

Let us show $(M,w) \models \mathbf{Ir}_i(\mathbf{WF}(\mathcal{A}))$ for all $\mathcal{A}$. Let $\mathcal{A}_k = \{A_k(s_k)\}_{s_k \in S_k}, k=1,2$ be given. Let $E_k = \{s_k \in S_k : (M,w) \models A_k(s_k)\}$ for $k=1,2$. First, notice, using (32), that if $(M,w) \models \neg[\mathrm{N}_1(\mathcal{A}) \wedge \mathrm{N}_2(\mathcal{A})]$, then $(M,w) \models WF_i(\mathcal{A})$. Thus, we can assume that $(M,w) \models \mathrm{N}_1(\mathcal{A}) \wedge \mathrm{N}_2(\mathcal{A})$. Using $\mathrm{N0}_1(\mathcal{A}) \wedge \mathrm{N0}_2(\mathcal{A})$, we have, for any $(s_1;s_2) \in S$, $(M,w) \models A_1(s_1) \wedge A_2(s_2) \supset \mathrm{bst}_1(s_1;s_2) \wedge \mathrm{bst}_2(s_2;s_1)$, i.e., $E_1 \times E_2 \subseteq E(G)$. Consider two cases.

(i) Suppose that $E_1 \times E_2 \subseteq F$. Then, by (31), for $k=1,2$, $(M,w) \models \wedge_{s_k \in S_k}[A_k(s_k) \supset \mathrm{I}_k(s_k)]$, and hence $(M,w) \models WF_i(\mathcal{A})$.

(ii) Suppose that $E_1 \times E_2 - F \neq \emptyset$. Because F is a subsolution, it is maximal having the form of $F = F_1 \times F_2$. Also by $E_1 \times E_2 \subseteq E(G)$, we have $F - E \neq \emptyset$. Let $(s_1^*, s_2^*) \in F - E$. Then, $(M,w) \models [\mathrm{I}_1(s_1^*) \wedge \mathrm{I}_2(s_2^*)] \wedge \neg[A_1(s_1^*) \wedge A_2(s_2^*)]$ and hence for $i=1,2$, $(M,w) \models \neg[\mathrm{I}_i(s_i^*) \wedge \mathbf{B}_j(\mathrm{I}_j(s_j^*)) \supset A_i(s_i^*) \wedge \mathbf{B}_j(A_j^*(s_j))]$. Thus, $(M,w) \models WF_i(\mathcal{A})$ for $i=1,2$.■

Proof of Theorem 5.1: Let G be an uninterchangeable game, and let F, F' be two subsolutions with $(s_i;s_j) \in F$ but $(s_i;s_j) \notin F'$. By Lemma 5.1, there are two models M and M' so that (28) and (29), respectively, for F and F'. Hence, $(M,w) \models \mathbf{B}_i(\mathrm{I}_i(s_i))$ but $(M',w') \not\models \mathbf{B}_i(\mathrm{I}_i(s_i))$. By soundness for IR^2, we have $\Delta_i \nvdash \neg\mathbf{B}_i(\mathrm{I}_i(s_i))$ and $\Delta_i \nvdash \mathbf{B}_i(\mathrm{I}_i(s_i))$.■

Since the model given in Lemma 5.1 has a single world, it is a model for Axioms T, 4 and 5. Hence, Theorem 5.1 holds for IR^2 with those axioms. In the following proof, we use Theorem 5.1 holds for $\mathrm{IR}^2(\mathrm{T})$.

Proof of Theorem 5.2. Suppose that there is a game formula A such that (22) holds in the logic IR^2; *a fortiori,* (22) holds for $\mathrm{IR}^2(\mathrm{T})$.Theorem 3.2 claims that in $\mathrm{IR}^2(\mathrm{T})$, $\mathbf{Ir}_i(\mathbf{g}) \vdash \mathbf{B}_i(A)$ or $\mathbf{Ir}_i(\mathbf{g}) \vdash \mathbf{B}_i(\neg A)$. This and the supposition imply $\Delta_i \vdash \mathbf{B}_i(\mathrm{I}_i(s_i))$ or $\Delta_i \vdash \mathbf{B}_i(\neg\mathrm{I}_i(s_i))$ in $\mathrm{IR}^2(\mathrm{T})$. This is impossible since Theorem 5.1 holds for $\mathrm{IR}^2(\mathrm{T})$.■

Proof of Theorem 5.3: When $s_i \in \hat{F}_i$, we have $\mathbf{Ir}_i^o(\mathbf{g}) \vdash A_i^{**}(s_i)$, which implies $\mathbf{Ir}_i^o(\mathbf{g}) \vdash \mathrm{I}_i(s_i) \supset A_i^{**}(s_i)$. In the other cases, by Lemma 4.1.(2), $\mathbf{Ir}_i^o(\mathbf{N}) \vdash \mathrm{I}_i(s_i) \supset C_i^*(s_i)$. Thus,

$$\mathbf{Ir}_i^o(\mathbf{g}), \mathbf{Ir}_i^o(\mathbf{N}) \vdash \ \mathrm{I}_i(s_i) \supset C_i^*(s_i) \text{ for all } s_i \in S_i. \tag{33}$$

Now, consider the converse of (33).

We modify Lemma 4.3 as follows: for any $(s_i;s_j) \in S$,

(0^*): $\mathbf{Ir}_i^o(\mathbf{g}), \mathbf{Ir}_i^o(\mathbf{N}) \vdash C_i^*(s_i) \wedge \mathbf{B}_j(C_j^*(s_j)) \supset \mathrm{bst}_i(s_i;s_j)$.

(1^*): $\mathbf{Ir}_i^o(\mathbf{g}), \mathbf{Ir}_i^o(\mathbf{N}) \vdash C_i^*(s_i) \supset \vee_{t_j \in S_j}\mathbf{B}_j(C_j^*(t_j))$.

(2^*): $\mathbf{Ir}_i^o(\mathbf{N}) \vdash C_i^*(s_i) \supset \mathbf{B}_j\mathbf{B}_i(C_i^*(s_i))$.

(0^*): If $C_i^*(s_i) = A_i^*(s_i)$ or $C_j^*(s_j) = A_j^*(s_j)$, then $\mathbf{Ir}_i^o(\mathbf{g}) \vdash \neg C_i^*(s_i)$ or $\mathbf{Ir}_i^o(\mathbf{g}) \vdash \mathbf{B}_j(\neg C_j^*(s_j))$; so, the assertion holds. Let $C_i^*(s_k) = A_i^{**}(s_i)$ and $C_j^*(s_j) = A_j^{**}(s_j)$. So, we have $\mathbf{Ir}_i^o(\mathbf{g}) \vdash \mathrm{bst}_i(s_i;s_j)$; so, we have the assertion. Let $C_i^*(s_k) = A_i^{**}(s_i)$ and $C_j^*(s_j) = \mathrm{I}_j(s_j)$. Then, for any $k = 1,...,l$, $(s_i;t_j) \in F^k$ for some t_j, and also, for some k_0, $(s_j;t_i) \in F^{k_0}$ for some t_j. Hence, we have $(s_i;s_j) \in F^{k_0}$, i.e., $(s_i;s_j)$ is a Nash equilibrium. Hence, $\mathbf{Ir}_i^o(\mathbf{g}) \vdash \mathrm{bst}_i(s_i;s_j)$. The case where $C_i^*(s_k) = \mathrm{I}_i(s_i)$ and $C_j^*(s_j) = A_j^{**}(s_j)$ is similar.

(1^*): First, let $C_i^*(s_i) = \mathrm{I}_i(s_i)$. By $\mathrm{N1}_i$, $\vdash C_i^*(s_i) \supset \vee_{t_j \in S_j}\mathbf{B}_j(\mathrm{I}_j(t_j))$. Then, since $\mathbf{Ir}_i^o(\mathbf{g}), \mathbf{Ir}_i^o(\mathbf{N}) \vdash$

$\mathbf{Ir}_j(\mathbf{g}) \wedge \mathbf{Ir}_j(\mathbf{N})$ by (6), we use (33) for j and get $\mathbf{Ir}_i^o(\mathbf{g}), \mathbf{Ir}_i^o(\mathbf{N}) \vdash \vee_{t_j \in S_j} \mathbf{B}_j(\mathrm{I}_j(t_j)) \supset \vee_{t_j \in S_j} \mathbf{B}_j(C_j^*(t_j))$. Thus, $\mathbf{Ir}_i^o(\mathbf{g}), \mathbf{Ir}_i^o(\mathbf{N}) \vdash C_i^*(s_i) \supset \vee_{t_j \in S_j} \mathbf{B}_j(C_j^*(t_j))$. Second, let $C_i^*(s_i) = A_i^*(s_i)$. Then, $\mathbf{Ir}_i^o(\mathbf{g}) \vdash \neg C_i^*(s_i)$, and hence, $\mathbf{Ir}_i^o(\mathbf{g}) \vdash C_i^*(s_i) \supset \vee_{t_j \in S_j} \mathbf{B}_j(C_j^*(t_j))$. Third, let $C_i^*(s_i) = A_i^{**}(s_i)$. Let $s_j \in \hat{F}_j$. Then, since $\vdash \mathbf{Ir}_i^o(\mathrm{bst}_i(s_i; s_j);\mathrm{bst}_j(s_j; s_i)) \supset \mathbf{Ir}_j(\mathrm{bst}_j(s_j; s_i);\mathrm{bst}_i(s_i; s_j))$ by (6), we have $\vdash C_i^*(s_i) \supset \vee_{t_j \in \hat{F}_j} \mathbf{B}_j(C_j^*(t_j))$. Then, $\vdash C_i^*(s_i) \supset [\vee_{t_j \in \hat{F}_j} \mathbf{B}_j(C_j^*(t_j))] \vee [\vee_{t_j \in S_j - \hat{F}_j} \mathbf{B}_j(C_j^*(t_j))]$, equivalently, $\vdash C_i^*(s_i) \supset \vee_{t_j \in S_j} \mathbf{B}_j(C_j^*(t_j))$.

(2*): If $C_i^*(s_i) = A_i^*(s_i)$, we have $\vdash C_i^*(s_i) \supset \mathbf{B}_j\mathbf{B}_i(C_i^*(s_i))$ by Lemma 4.3.(2). The case for $C_i^*(s_i) = A_i^{**}(s_i)$ is similar. If $C_i^*(s_i) = \mathrm{I}_i(s_i)$, then $\vdash C_i^*(s_i) \supset \mathbf{B}_j\mathbf{B}_i(C_i^*(s_i))$ by N2$_i$.■

The above three statements imply $\mathbf{Ir}_i^o(\mathbf{g}), \mathbf{Ir}_i^o(\mathbf{N}) \vdash \mathrm{N}_i(\mathcal{C}^*) \wedge \mathbf{B}_j(\mathrm{N}_j(\mathcal{C}^*))$, and also, by (33), we have $\mathbf{Ir}_i^o(\mathbf{g}), \mathbf{Ir}_i^o(\mathbf{N}) \vdash \wedge_{s \in S}\langle \mathrm{I}_i(s_i) \wedge\ \mathbf{B}_j(\mathrm{I}_j(s_j)) \supset C_i^*(s_i) \wedge \mathbf{B}_j(C_j^*(s_j))\rangle$. Then, we using $\mathbf{Ir}_i^o(\mathbf{WF}(\mathcal{C}^*))$, we have $\mathbf{Ir}_i^o(\mathbf{g})$, $\mathbf{Ir}_i^o(\mathbf{N}), \mathbf{Ir}_i^o(\mathbf{WF}(\mathcal{C}^*)) \vdash C^*(s_i) \supset \mathrm{I}_i(s_i)$.■

Proof of Theorem 5.4: (Only-if): Suppose $(s_i; s_j) \notin \hat{F}$ for any $s_j \in S_j$. Let s_i be not a Nash strategy. Then, $\Delta_i \vdash \mathbf{B}_i(\neg \mathrm{I}_i(s_i))$ by (20); so $\Delta_i \vdash \neg \mathbf{B}_i(\mathrm{I}_i(s_i))$ by Axiom D. Since Δ_i is consistent by Lemma 4.2, we have $\Delta_i \nvdash \mathbf{B}_i(\mathrm{I}_i(s_i))$. Let s_i be a Nash strategy. Then, $s_i \notin F_i^l$ for some subsolution $F_1^l \times F_2^l$. Thus, $\Delta_i \nvdash \mathbf{B}_i(\mathrm{I}_i(s_i))$ by (23).

(If): If $(s_i; s_j) \in \hat{F}$ for some s_j, then $\mathbf{Ir}_i^o(\mathbf{g}) \vdash A^{**}(s_i)$. Hence, $\Delta_i^o \vdash \mathrm{I}_i(s_i)$ by Theorem 5.3, which implies $\Delta_i \vdash \mathbf{B}_i(\mathrm{I}_i(s_i))$.■

6 Conclusions

We have considered prediction/decision making by player i in a finite 2-person game G. His decision criterion constitutes of the three axioms, $\mathrm{N}_i = \mathrm{N0}_i \wedge \mathrm{N1}_i \wedge \mathrm{N2}_i$, which are described in the mind of player i, and we require the same for the other player j. Therefore, player i is led to an infinite regress consisting of N_i and N_j. This infinite regress is captured by $\mathbf{Ir}_i(\mathrm{N}_i;\mathrm{N}_j)$ in the fixed-point extension IR^2 of the epistemic logic KD^2. We adopted this $\mathbf{Ir}_i(\mathbf{N}) = \mathbf{Ir}_i(\mathrm{N}_i;\mathrm{N}_j)$ and the additional infinite regresses, $\mathbf{Ir}_i(\mathbf{WF})$ and $\mathbf{Ir}_i(\mathbf{g})$. For an interchangeable game G, the belief set $\Delta_i = \{\mathbf{Ir}_i(\mathbf{g}), \mathbf{Ir}_i(\mathbf{N})\} \cup \mathbf{Ir}_i(\mathbf{WF})$ determine $\mathrm{I}_i(s_i)$ to the some specific formula $A^*(s_i)$, while the situation for an uninterchangeable G is entirely different. Here, we discuss various relevant points for our results on decidability and undecidability.

Positive, negative decisions, and undecidable: Suppose that G has the interchangeable set of Nash equilibria. Our decidability result states that player i finds his Nash strategy to be a possible decision, and disproves any non-Nash strategy as a negative decision. Player i may find multiple possible decisions, but can use any for his play. Our theory is silent for this choice.

Suppose that G has no Nash equilibria. The decidability result states that player i denies any strategies as negative decisions. Then, the negative decisions may lead player i to a different decision criterion such as a *default* criterion, e.g., the first strategy should be chosen, to the necessity of communication.

On the other hand, when G has multiple subsolutions, we presented the undecidability result that player i cannot find any positive decision, unless the entire subsolutions has a nonempty intersection. In this case, he can reach neither a positive nor a negative decision. Then, he cannot go to a new criterion, since he himself does not notice undecidability.

A **way out?** We may regard communication between the two players as a way out. Technically, if they can communicate with each other, a specifying a subsolution to be chosen would resolve undecidability. Again, difficulty is that player i does not notice this necessity.

Two independent minds and a discord in $\mathbf{Ir}_i(\mathbf{g})$: Gödel's theorem is caused by the self-referential structure of Peano Arithmetic. That is, the entire theory of Peano Arithmetic can be described inside the theory; the contradiction-freeness of Peano Arithmetic is one important example. Our framework includes also a self-referential structure; the infinite regress operator $\mathbf{Ir}_i(\cdot;\cdot)$ describes $\mathbf{Ir}_j(\cdot;\cdot)$, and *vice versa* in the logic IR^2. This gives an environment for our undecidability, but does not directly generate it. For example, the prediction/decision criterion $\{\mathbf{Ir}_i(\mathbf{N})\} \cup \mathbf{Ir}_i(\mathbf{WF})$ treats the two minds of the players equally in that each is embedded into the other. In fact, a discord between the two players is included only in the infinite regress of the game $\mathbf{Ir}_i(\mathbf{g})$. This discord is the real source for our undecidability.

The self-referential involved in our framework serves an environment for undecidability, since it gives simultaneous decisions and predictions for the two players. In the end of this section, we discuss other environments where non-simultaneous decisions are made and where we have only decidability results.

Johansen's [10] argument: He gave the following four postulates for prediction/decision making and asserted that the Nash noncooperative solution could be derived from them for solvable games. He assumed (p.435) that the game has the unique Nash equilibrium.

Postulate J1 (Closed world): A player makes his decision $s_i \in S_i$ on the basis of, and only on the basis of information concerning the action possibility sets of two players S_1, S_2 and their payoff functions h_1, h_2.

Postulate J2 (Symmetry in rationality): In choosing his own decision, a player assumes that the other is rational in the same way as he himself is rational.

Postulate J3 (Predictability): If any[7] decision is a rational decision to make for an individual player, then this decision can be correctly predicted by the other player.

Postulate J4 (Optimization against "for all" predictions): Being able to predict the actions to be taken by the other player, a player's own decision maximizes his payoff function corresponding to the predicted actions of the other player.

Those are connected to our requirements $\mathrm{N0}_i \wedge \mathrm{N1}_i \wedge \mathrm{N2}_i$ and $\mathrm{N0}_j \wedge \mathrm{N1}_j \wedge \mathrm{N2}_j$. J1 requires player i's prediction/decision criterion to be described by game formulae. J3 corresponds to $\mathrm{N2}_i$ and $\mathrm{N2}_j$, and J4 to $\mathrm{N0}_i$ and $\mathrm{N0}_j$. Here, J2 should be interpreted as symmetry between $\mathrm{N0}_i \wedge \mathrm{N1}_i \wedge \mathrm{N2}_i$ and $\mathrm{N0}_j \wedge \mathrm{N1}_j \wedge \mathrm{N2}_j$ together with the symmetric treatment of two minds in IR^2. Complete symmetry is obtained in terms of infinite regresses $\{\mathbf{Ir}_i(\mathbf{N})\} \cup \mathbf{Ir}_i(\mathbf{WF})$, while keeping the identities of two minds. Once $\mathbf{Ir}_i(\mathbf{g})$ is introduced, it may contain some discord, which may generate undecidability. Johansen himself did not discuss this part at all.

Other undecidability in game theory: The undecidability result given in Kaneko-Nagashima [11] takes the same form as our undecidability[8]. They gave a 3-person game having a unique Nash equilibrium in mixed strategies. It is assumed that the game structure and real number theory Φ_{rcf} (real closed field theory) are common knowledge among the players. They proved the provability of $\mathbf{C}(\exists x \mathrm{Nash}(x))$ from their common knowledge of G and Φ_{rcf}. However, from the same common knowledge assumption, neither $\exists x \mathbf{C}(\mathrm{Nash}(x))$ nor $\neg \exists x \mathbf{C}(\mathrm{Nash}(x))$ is provable. That is, the players commonly know the abstract existence of a Nash equilibrium, but do not

[7] This "any" was "some" in Johansen's orginal Posutlate 3. According to logic, this should be "any". However, this is expressed as "some" by many scientists (even mathematicians).

[8] There are some literature on uncomputability on optimal strategies in a simple extensive game

find a concrete one; hence they cannot play the specific Nash equilibrium strategy.

This is related to neither a self-referential structure nor the interdependence of the situation. It is caused by the lack of the names of irrational numbers such as $\sqrt{51}$ in their language, which is involved in the Nash equilibrium in the 3-person game with rational payoffs[9]. The main reason for this difficulty is to give a name to a concept, but not the self-referential structure.

Other solution concepts in game theory: Theorem 4.1 appears related to Aumann-Brandenburger [2] in that "Nash equilibrium" is derived there in a game model. It would be difficult make a direct comparison with their model in that it is a game model of decision making following the Bayesian-game theory tradition, but not a model in the sense of logic. As remarked, our target is a possible decision but Nash equilibrium is a realization of it. Anyhow, since it is a single model, it is incapable of talking about undecidability like ours. Also, it is worth mentioning that a solution function there describing a decision is single-valued, while we consider a possible decision, to be interpreted as a set-(possibly empty)-valued function. Therefore, the Nash solution theory was not questioned in [2].

The game theory literature has various "solution concepts" other than the Nash solution theory. As far as we have checked, confining to finite games with pure strategies, there are no solution concepts for undecidability other than the Nash solution theory.

For example, the theory of "rationalizable strategies" (cf., Osborne-Rubinstein [18]) can be formulated by a similar axiomatization to $\mathrm{Na}_1-\mathrm{Na}_2$, except that "for all predictions" is replaced by "some prediction". Then, a variant of $\mathrm{N0}_1$ and $\mathrm{N0}_2$ can axiomatize "rationalizable strategies". The full axiomatization including beliefs can be done in the infinitary logic in Hu, *at al.* [9][10]. Here, when an appropriate belief set of payoffs is given, we have the decidability.

Dominant strategy criterion: Let us see the *dominant strategy criterion.* In addition to $\mathrm{N0}_i$ as the basic axiom, and we assume the following axiom for predictions, instead of $\mathrm{N1}_i$ and $\mathrm{N2}_i$:

Dm_i **(Giving up prediction)**: $\wedge_{s_j \in S_j}[\mathbf{B}_j(\mathrm{I}_j(s_j))]$.

It states that player i gives up predicting j's decision by accepting any strategy for j as a possible decision. Player i's thinking is already closed in $\mathrm{N0}_i$ and Dm_i, i.e., $\mathbf{B}_i(\mathrm{N0}_i \wedge \mathrm{Dm}_i)$.

We consider the formula $\mathrm{dm}_i(s_i) := \wedge_{t_j \in S_j} \mathrm{bst}_i(s_i; t_j)$ expressing "s_i is a dominant strategy". Then, $\mathbf{B}_i(\mathrm{N0}_i \wedge \mathrm{Dm}_i) \vdash \mathbf{B}_i(\mathrm{I}_i(s_i)) \supset \mathbf{B}_i(\mathrm{dm}_i(s_i))$ for all $s_i \in S_i$. The converse can be formulated in the similar manner as $\mathrm{WF}_i(\mathcal{A}_i)$ in Section 4.2: Let $\mathbf{WF}_i^{dm} = \{\mathrm{Dm}_i(\mathcal{A}) \supset \wedge_{s \in S}[A_i(s_i) \supset \mathrm{I}_i(s_i)] : \mathcal{A}_i = \{A_i(s_i)\}_{s_i \in S_i}\}$. Then, we have the following theorem:

$$\mathbf{B}_i(\mathrm{N0}_i \wedge \mathrm{Dm}_i), \mathbf{B}_i(\mathbf{WF}_i^{dm}) \vdash \mathbf{B}_i(\mathrm{I}_i(s_i) \equiv \mathrm{dm}_i(s_i)). \tag{34}$$

This has two important points: First, the epistemic depth for this result is 1; no interdependency between the two players are involved. Second, Theorem 3.1 (decidability) implies also decidability.

The result (34) is extended in various manners: We may specify some strategies only for predictions. Or, player i assumes that player j follows Dm_j; then interpersonal interdependence

[9] Classical game theory for the 2-person case can be done only in rational numbers. For the 3-person case, any algebraic real numbers in $[0, 1]$ are involved as some mixed strategy equilibria Nevertheless, if they are assumed to be expressed by constants, which is possible, we can avoid the undecidability in [11].

[10] A fixed-point logic approach is also possible, but it needs a specific formulation, and is more cumbersome than IR^2 of this paper. The infinitary logic approach gives a unified way for the Nash theory and rationalizability theory.

of degree 2 is required. Still, we can formulate those criteria without conceptual difficulties. For those case, we have decidability as far as the beliefs of payoffs are given in an appropriate manner. More generally, if we start with the same argument in a finite but repeated way, we have only decidability. We, however, emphasize that those extensions do not give a completely symmetric environment. In this sense, the self-referential structure for the two players is crucial for our undecidability result.

References

[1] Aumann, R. J. (1976), Agreeing to Disagree, *Annals of Statistics* 4, 1236–1239.

[2] Aumann, R. J., and A. Brandenburger, (1995), Epistemic Conditions for Nash Equilibrium, *Econometrica* 63, 1161-1180.

[3] Boolos, G., (1979), *The Unprovability of Consistency*, Cambridge University Press, Cambridge.

[4] Brandenburger, A., (2014), *The Language of Game Theory*, World Scientific, London.

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

[6] Heifetz, A., (1999), Iterative and Fixed Point Common Belief, *Journal of Philosophical Logic* 28, 61-79.

[7] Hu, T., and M. Kaneko (2012), Critical Comparisons between the Nash Noncooperative Theory and Rationalizability, *Logic and Interactive Rationality Yearbook 2012*, Vol.II, eds. Z. Christo, *et al.* 203-226, http://www.illc.uva.nl/dg/?page_id=78

[8] Hu, T., and M. Kaneko (2014), Infinite Regress Logic.

[9] Hu, T., M. Kaneko, and N.-Y. Suzuki, (2014), Small Infinitary Epistemic Logics and Some Fixed-Point Logics, to appear in August.

[10] Johansen, L., (1982), On the Status of the Nash Type of Noncooperative Equilibrium in Economic Theory, *Scand. J. of Economics* 84, 421-441.

[11] Kaneko, M., and T. Nagashima, (1996), Game logic and its applications I, *Studia Logica* 57, 325–354.

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

[13] Kline, J. J., (2013), Evaluations of epistemic components for resolving the muddy children puzzle, *Economic Theory* 53, 61-84.

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

[15] Meyer, J.-J. Ch., van der Hoek, W., (1995), *Epistemic logic for AI and computer science.* Cambridge.

[16] Mendelson, E., (1988), *Introduction to Mathematical Logic*, Wadsworh, Monterey.

[17] Nash, J. F., (1951), Non-cooperative Games, *Annals of Mathematics* 54, 286-295.

[18] Osborne, M., and A. Rubinstein, (1994), *A Course in Game Theory*, MIT Press, Cambridge.

[19] Perea, A., (2012), *Epistemic Game Theory: Reasoning and Choice,* Cambridge University Press, Cambridge.

[20] Suzuki, N.-Y., (2013), Semantics for intuitionistic epistemic logics of shallow depths for game theory, *Economic Theory* 53, 85-110.

[21] Van Benthem, *Logic in Games*, Institute for Logic, Language and Computation.

[22] Van Benthem, J., E. Pacuit, and O. Roy, (2011), Toward a Theory of a Play: A Logical Perspective on Games and Interaction, *Games* 2, 52-86.