Game Theoretic Decidability and Undecidability[*,†]

Tai-Wei Hu[‡] and Mamoru Kaneko[§]

07 October 2014; this version, 29 January 2015 (Incom_01-29-15-HK)

Abstract

We study the possibility of prediction/decision making in a finite 2–person game with pure strategies, following the Nash-Johansen noncooperative solution theory. We adopt the infinite-regress logic EIR^2 (a fixed-point extension) of the epistemic logic KD^2 to capture individual decision making from the viewpoint of logical inference. In the logic EIR^2, prediction/decision making is described by the belief set $\Delta_i(\mathbf{g})$ for player i, where $\mathbf{g}$ specifies a game. Our results on prediction/decision making differ between solvable and unsolvable games. For the former, we show that player i can decide whether each of his strategies is a final decision or not. For the latter, we obtain undecidability, i.e., he can neither decide some strategy to be a possible decision nor disprove it. Thus, the theory $(\mathrm{EIR}^2; \Delta_i(\mathbf{g}))$ is incomplete in the sense of Gödel's incompleteness theorem for an unsolvable game $\mathbf{g}$. This result is related to "self-referential", but its main source is a discord generated by interdependence of payoffs and independent prediction/decision making.

Key words: Prediction/decision making, Infinite regress, Game theoretic decidability, Undecidability, Incompleteness, Nash solution, Subsolution

1 Introduction

Logical inference is an engine for decision making in games with two or more players. Although game theory has studied decision making extensively, logical inference is kept informal. To study the inference aspect of the decision making process in games, we adopt a formal system of epistemic logic, the *epistemic infinite-regress logic* EIR^2, developed in Hu-Keneko [7]. It is a fixed-point extension of the (propositional) epistemic logic KD^2. We focus on the 2-person case for simplicity. Because of interdependence of players, prediction making is also required, and our logic allows us to model prediction making based on logical inference. Our approach also emphasizes players' independent thinking, and this emphasis guides our formulation of EIR^2. Our approach is coherent with Nash [17] and Johansen [9], who gave the noncooperative theory of prediction/decision making in a non-formalized manner.

We prove game theoretic decidability and undecidability, depending upon whether a game has the interchangeable set of Nash equilibria. The former result states that a player can reach

[*]The authors thank Nobu-Yuki Suzuki and Johan van Benthem for valuable suggestions on this research.

[†]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, mkanekoepi@waseda.jp

a positive or a negative decision for each strategy, while the latter states that for some strategy, he cannot reach either a positive or a negative decision. Our approach takes various perspectives different from the standard literature of game theory as well as that of epistemic logic. We lay out those perspectives in the following.

Fixed-point extension of KD^2: In game situations, prediction/decision making naturally leads to an infinite regress of beliefs. This regress begins subjectively in the mind of player i in his prediction making, which requires him to simulate the other player's mind, and in such simulation another layer of interpersonal thinking is required; the regress would go *ad infinitum* unless we stop it at an arbitrary layer. We adopt the fixed-point extension EIR^2 of KD^2, to capture this infinite regress[1]. An infinite number of imaginary players are involved in this regress, and their scopes are distinguished in the logic EIR^2.

As the concept of infinite regress of beliefs is closely related to common knowledge, the logic EIR^2 is also related to the common knowledge logic CKL (cf., Fagin, *et al.* [4], and Meyer-van der Hoek [15])[2]. In fact, when we add Axiom T(truthfulness) to EIR^2, infinite regress collapse to common knowledge, and the resulting logic EIR^2(T) is equivalent to CKL. Without Axiom T, however, EIR^2 can capture mutual subjectivity, which is not allowed in CKL. For this reason, even though some results in this paper are sharper in EIR^2(T) than in EIR^2, we take the latter as the basic system[3].

Proof theory and model theory: Because of our focus on logical inference in prediction/decision making, we use a proof-theoretic system, which allows us to formulate a player's reasoning process explicitly. This approach is in sharp contrast with most models in epistemic game theory[4], which describe possible mental states in a single (semantic) model[5]. We also use model theory (here, Kripke semantics) as a technical support, which is connected to our formal system via the soundness/completeness theorem (see Hu-Kaneko [7]). By soundness/completeness for EIR^2, we can use Kripke models to evaluate provability via validity or via finding a counter-model. In particular, the soundness part will be used to prove our game theoretic undecidability result.

Basic beliefs as non-logical axioms: We formulate basic beliefs as axioms for a player's prediction/decision making in the logic EIR^2. Those basic beliefs include his understanding of the game and prediction/decision criterion. The derivation from his beliefs to a possible final decision is expressed as

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

[1] Alternatively, we can adopt an infinitary logic. Hu *et al.* [8] discusses relationships between the logic EIR^2 and its infinitary counterpart.

[2] It is also related to "common belief" (cf., Heifetz [5]), but here, we compare only common knowledge with our concept of infinite regress.

[3] Some reader may recall a general theory of fixed-point logics, called the μ-calculus, from the semantic side (cf., Venema [23] and van Benthem [20], Chap.22). We adopt the logic EIR^2 to capture epistemic infinite regresses arising from game theoretic prediction/decision making in terms of fixed-point arguments, instead of discussing fixed-point properties themselves. Thus, we consider a particular type of fixed-points in EIR^2. In this sense, it is more related to the common knowledge logic CKL mentioned above.

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

[5] 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], the various papers in Brandenbuger [3], and van Benthem [21]. Some exceptions are Kaneko-Nagashima [10], Kline [14], and Suzuki [19], where the proof-theoretic standpoint is taken.

That is, player i has basic beliefs Γ_i^o in his mind, and derives $\mathrm{I}_i(s_i)$ as a logical conclusion. His beliefs then recommend s_i as a possible final decision. The negative decision is expressed as $\mathbf{B}_i(\Gamma_i^o) \vdash \mathbf{B}_i(\neg \mathrm{I}_i(s_i))$; his beliefs recommend him not to take s_i. Although (1) is expressed from the analyst's viewpoint, we intend to model these derivations as occurring in player i's mind. This is justified by Lemma 2.5, which implies that in EIR^2, $\mathbf{B}_i(\Gamma_i^o) \vdash \mathbf{B}_i(\mathrm{I}_i(s_i))$ (respectively, $\mathbf{B}_i(\Gamma_i^o) \vdash \mathbf{B}_i(\neg \mathrm{I}_i(s_i))$) is equivalent to $\Gamma_i^o \vdash \mathrm{I}_i(s_i)$ ($\Gamma_i^o \vdash \neg \mathrm{I}_i(s_i)$). The choice of the base logic KD^2 is essential for this equivalence.

Game theoretic concepts: We only consider finite 2-person strategic games with pure strategies, which are sufficiently rich to obtain both decidability and undecidability results. The characterization of games with decidability/undecidability corresponds to the solvability/unsolvability requirement in Nash [17]. Solvability captures players' independence in *ex ante* prediction/decision making, though Nash did not distinguish predictions from decisions. Johansen [9], making this distinction, discussed Nash's theory in a philosophical manner. As our axioms for prediction/decision making formalize Nash's and Johansen's arguments in the logic EIR^2, the resulting system is called the *formalized Nash (-Johansen) theory*.

This theory features a symmetric treatment of each player's thinking about the game as well as his decision making, which leads to an infinite regress of interpersonal beliefs. Although Nash equilibrium naturally arises in deriving possible final decisions, our main focus is on the prediction/decision making process, in particular, on whether the process reaches a final conclusion or not. Thus, we treat Nash equilibrium as an auxiliary concept in the formalized Nash theory[6].

Axiomatic formulation of prediction/decision making: The Nash theory could be described in a nonformalized language: 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 is the set of possible final decisions for i, and E_j i's prediction about j's decisions; completion of the meaning stated in Na_i requires Na_j, and vice versa. Hu-Kaneko [6] proved that this pair characterizes the Nash theory in a non-formalized language, which is given as Proposition 3.1.

We formalize Na_i as Axiom $\mathrm{N0}_i$ in the logic EIR^2 taking players' beliefs explicitly into account. In the formalized language, this requires us to impose two additional axioms, $\mathrm{N1}_i$ and $\mathrm{N2}_i$; the first appears because of explicit consideration of beliefs and the second corresponds to the non-emptiness requirement. We take those three axioms, $\mathrm{N0}_i$,$\mathrm{N1}_i$, $\mathrm{N2}_i$ as the start for player i's prediction/decision making; those are in the scope of player i's mind, expressed as $\mathbf{B}_i(\mathrm{N012}_i) := \mathbf{B}_i(\mathrm{N0}_i \wedge \mathrm{N1}_i \wedge \mathrm{N2}_i)$. As Na_i and Na_j require each other, $\mathbf{B}_i(\mathrm{N012}_i)$ needs the counterpart for player j. Because the players' beliefs are now explicit, the counterpart is expressed as $\mathbf{B}_i\mathbf{B}_j(\mathrm{N012}_j)$, where $\mathrm{N012}_j$ is the same as $\mathrm{N012}_i$ by the replacement of i with j. Thus, for player i to predict about j's decision, i thinks about j's beliefs. For the same reason, $\mathbf{B}_i\mathbf{B}_j(\mathrm{N012}_j)$ requires $\mathbf{B}_i\mathbf{B}_j\mathbf{B}_i(\mathrm{N012}_i)$, and so on. Thus, to complete prediction making, player i would meet an infinite regress of interpersonal beliefs:

$$\mathbf{B}_i(\mathrm{N012}_i),\ \mathbf{B}_i\mathbf{B}_j(\mathrm{N012}_j),\ \mathbf{B}_i\mathbf{B}_j\mathbf{B}_i(\mathrm{N012}_i), ... \tag{2}$$

In the logic EIR^2, this infinite sequence is captured by the fixed-point operator, $\mathbf{Ir}_i(\mathrm{N012}_i;\mathrm{N012}_j)$.

[6] Aumann-Brandenburger [1] gave a model and axioms for players' decision making in a game. They derived Nash equilibrium from their axioms. We also consider player's prediction/decision making, and it is naturally related to the Nash solution theory. Nash equilibrium is an auxiliary building block of our theory.

The infinite sequence (2), *a fortiori,* $\mathbf{Ir}_i(\mathrm{N012}_i;\mathrm{N012}_j)$, has a self-referential structure: The sequence itself occurs with the scope of $\mathbf{B}_i(\cdot)$, the counterpart for player j is with the scope of $\mathbf{B}_i\mathbf{B}_j(\cdot)$, and (2) again occurs again with $\mathbf{B}_i\mathbf{B}_j\mathbf{B}_i(\cdot)$, and so on. This self-referential structure is crucial for our undecidability result.

The infinite regress, $\mathbf{Ir}_i(\mathrm{N012}_i;\mathrm{N012}_j)$, is our basic postulate for prediction/decision making, but it only provides a necessary condition for possible final decisions. For a full determination of a possible decision, we need another axiom (schema), $\mathbf{Ir}_i(\mathbf{WF})$, which would choose logically weakest formulae satisfying the property described by $\mathbf{Ir}_i(\mathrm{N012}_i;\mathrm{N012}_j)$.

Formalized Nash theory: The set of beliefs $\mathbf{Ir}_i(\mathrm{N012}_i;\mathrm{N012}_j), \mathbf{Ir}_i(\mathbf{WF})$ describes prediction/decision making without concrete information about the game being played. We formulate the basic beliefs of a game, including strategies and payoffs, by $\mathbf{Ir}_i(\mathbf{g}) := \mathbf{Ir}_i(g_i; g_j)$. This addition completes our postulates of player i's basic beliefs: $\Delta_i(\mathbf{g}) = \{\mathbf{Ir}_i(\mathbf{g}), \mathbf{Ir}_i(\mathrm{N012}_i;\mathrm{N012}_j)\}$ $\cup \mathbf{Ir}_i(\mathbf{WF})$, which plays the role of $\mathbf{B}_i(\Gamma_i^o)$ in (1). The pair $(\mathrm{EIR}^2; \Delta_i(\mathbf{g}))$ of the logic EIR^2 and player i's basic beliefs forms the *formalized Nash theory for the game* $\mathbf{g}$.

The literature of game theory tends to focus on the resulting outcome(s) from a solution/equilibrium theory. In our context, this focus can be stated as the following question:

(i): What decisions and predictions does $(\mathrm{EIR}^2; \Delta_i(\mathbf{g}))$ recommend?

This question presumes that the theory $(\mathrm{EIR}^2; \Delta_i(\mathbf{g}))$ has recommendations. However, we should ask the following question in the first place.

(ii): Does $(\mathrm{EIR}^2; \Delta_i(\mathbf{g}))$ recommend any decision?

In fact, the answer to the second question is related to Nash's [17] solvability condition.

We say that a game is *solvable* when the set of Nash equilibria is interchangeable, i.e., the set has a product structure. Here, we give three examples of games; two are solvable and the other is not. In Table 1.1, each player has three strategies, and his payoff is given in the matrix (the first and second entries are players 1's and 2's payoffs). The superscript NE stands for Nash equilibrium, defined in Section 3. Table 1.1 has a unique NE. Table 1.2, called the *battle of the sexes*, has two NE's; this is not solvable because the set is not a product set. Table 1.3, called the *matching pennies,* has the empty set of NE's. Tables 1.1 and 1.3 are solvable games.

Table 1.1

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

Table 1.2

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

Table 1.3

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

Positive, negative decisions, and undecidable: Our results give a complete answer to question (ii). For a solvable game, we have the decidability result: for *any* strategy s_i for player i,

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

Moreover, a strategy is positively recommended if and only if it is a component of a NE. For Table 1.1, the set of beliefs $\Delta_1(\mathbf{g})$ recommends player 1 to take $\mathbf{s}_{12}$ as a positive decision but not to take either $\mathbf{s}_{11}$ or $\mathbf{s}_{13}$. In Table 1.3, $\Delta_1(\mathbf{g})$ recommends all strategies as negative decisions.

We show that when $\mathbf{g}$ is not solvable as in Table 1.2, there is *some* strategy s_i for each player

i such that

$$\Delta_i(\mathbf{g}) \nvdash \mathbf{B}_i(\mathrm{I}_i(s_i)) \quad \text{and} \quad \Delta_i(\mathbf{g}) \nvdash \mathbf{B}_i(\neg \mathrm{I}_i(s_i)). \tag{4}$$

That is, player i cannot decide with the belief set $\Delta_i(\mathbf{g})$ whether s_i is a positive or negative decision. In Table 1.2, this holds for both strategies. This situation differs entirely from the case where $\Delta_i(\mathbf{g})$ gives negative recommendations for all strategies such as in Table 1.3; in the latter case, he may look for a different way to make his decisions, but in the former, i.e., (4), he may not be able to notice this undecidability itself, and get stuck in his decision making.

The undecidability result, (4), obtained under the formalized Nash theory, $(\mathrm{EIR}^2; \Delta_i(g))$, brings to light a new critical issue on ex ante decision-making that is not considered in the literature. While multiplicity is often recognized as an issue for prediction from the outside observer's perspective, here we show that undecidability can be a serious issue for prediction even from the *inside* players' perspectives.

Relations to Gödel's incompleteness theorem: The result (4) has the same form as Gödel's incompleteness theorem (cf., Mendelson [16]), noting that we can also prove a slightly stronger statement than the latter part of (4), $\Delta_i(\mathbf{g}) \nvdash \neg \mathbf{B}_i(\mathrm{I}_i(s_i))$. However, both interpretation and source for incompleteness differ. Gödel's theorem is about the Peano Arithmetic as the theory (CL; PA), where CL is the first-order predicate logic CL and PA is the set of axioms for natural numbers including the induction principle. Our theory is also given as the pair $(\mathrm{EIR}^2; \Delta_i(\mathbf{g}))$. Gödel's theorem is based on the self-referential structure. Although the self-referential structure is also important to obtain (4), the critical source is a discord included in the game g. Among the three components of $\Delta_i(\mathbf{g}) = \{\mathbf{Ir}_i(g), \mathbf{Ir}_i(\mathrm{N012}_i;\mathrm{N012}_j)\} \cup \mathbf{Ir}_i(\mathbf{WF})$,the second and third are completely symmetric between the two players, which form infinite regresses of themselves. The source for undecidability can then be attributed to the discord in $\mathbf{Ir}_i(\mathbf{g})$. A detailed comparison with Gödel's theorem will be given in Section 6.

The paper proceeds as follows: Section 2 formulates the logic EIR^2. Section 3 gives various game theoretic concepts. Section 4 gives three axioms for prediction/decision making, and the game theoretic decidability result for a solvable game. Section 5 presents the undecidability result for an unsolvable game. Section 6 gives concluding remarks.

2 The Epistemic Infinite Regress Logic EIR^2

We formulate the logic EIR^2 with the language for 2-person strategic games in Sections 2.1, 2.2, and give its semantics in Section 2.3. The language presumes the sets of strategies but this restriction is not essential for our argument.

2.1 Language

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

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

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

The atomic formula $\mathrm{Pr}_i(s; t)$ 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);[7]

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 stipulate that j refers to the other player than i. Player i's prediction about j's decision is expressed as $\mathbf{B}_j(\mathrm{I}_j(s_j))$, but this should occur in the scope of $\mathbf{B}_i(\cdot)$. 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 player i's subjective belief of the infinite regress of beliefs about A_i and A_j. 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 induction: (o) all atomic formulae are formulae; (i) if A, B are formulae, so are $(A \supset B)$, $(\neg A)$, $\mathbf{B}_i(A)$, $i = 1, 2$; (ii) if $\mathbf{A} = (A_1, A_2)$ is a pair of formulae, then $\mathbf{Ir}_i(\mathbf{A})$ is also a formula; and (iii) if Φ is a finite (nonempty) set of formulae, $(\wedge\Phi)$ and $(\vee\Phi)$ are formulae[8]. 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 $[,]$. We also write $\wedge\mathbf{B}_i(\Phi)$ for $\wedge\{\mathbf{B}_i(A) : A \in \Phi\}$, and etc.

2.2 Proof theory of EIR^2

We start with an explicit formulation of classical logic, which consists of 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 $A \in \Phi$; and **L5** $A \supset \vee\Phi$, where $A \in \Phi$;

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

The epistemic logic KD^2 is defined by adding, to classical logic, two epistemic axioms and one inference rule for the belief operators $\mathbf{B}_i(\cdot)$: for all formulae A, C, and for $i = 1, 2$,

K $\mathbf{B}_i(A \supset C) \supset (\mathbf{B}_i(A) \supset \mathbf{B}_i(C))$; and **D** $\neg\mathbf{B}_i(\neg A \wedge A)$;

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

Then, we have the *epistemic infinite regress logic* EIR^2, by adding one axiom (schema) and one inference rule for the infinite regress operators $\mathbf{Ir}_i(\cdot,\cdot)$: For $i = 1, 2$, and two pairs of formulae $\mathbf{A} = (A_1, A_2)$, $\mathbf{D} = (D_1, D_2)$,

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

[7] 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.

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

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:

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

The infinite sequence in (2) is a special case for (5). Rule IRI_i chooses $\mathbf{Ir}_i(\mathbf{A})$ as a logically weakest formula satisfying the property described in IRA_i, that is, if D_i enjoys it, then D_i implies $\mathbf{Ir}_i(\mathbf{A})$. The soundness/completeness result (Theorem 2.1) shows that $\mathbf{Ir}_i(\mathbf{A})$ captures faithfully the set in (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: (i) for each node $x \in X$, $\psi(x)$ is a formula attached to x; (ii) for each leaf x in $\langle X, <\rangle$, $\psi(x)$ is an instance of the axiom schemata; and (iii) 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.

As discussed in Section 1, non-logical axioms (beliefs for player i) play crucial roles in our study of prediction/decision making. Non-logical axioms are formulated as follows: 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[9].

The following results are basic to classical logic and/or KD^2 (cf., Kaneko [13]). 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)** $\vdash \mathbf{B}_i(\wedge\Phi) \equiv \wedge\mathbf{B}_i(\Phi)$.*

The following lemma enables us to consider the epistemic content of $\mathbf{Ir}_i(\mathbf{A})$, which is crucial in our game theoretical considerations in Sections 4 and 5.

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 see $\vdash \mathbf{B}_i(A_i \wedge \mathbf{Ir}_j(\mathbf{A})) \supset \mathbf{Ir}_i(\mathbf{A})$. 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 $\vdash \mathbf{B}_i(A_i) \wedge \mathbf{B}_i(\mathbf{Ir}_j(\mathbf{A})) \supset \mathbf{Ir}_i(\mathbf{A})$.

By the above result for j, $\vdash \mathbf{B}_i(D_j) \supset \mathbf{B}_i(\mathbf{Ir}_j(\mathbf{A}))$; so, $\vdash \mathbf{B}_i(A_i) \wedge \mathbf{B}_i(D_j) \supset \mathbf{B}_i(A_i) \wedge \mathbf{B}_i(\mathbf{Ir}_j(\mathbf{A}))$. Since $\vdash \mathbf{Ir}_i(\mathbf{A}) \supset \mathbf{B}_i(A_i) \wedge \mathbf{B}_i(D_j)$ by IRA_i, we have $\vdash \mathbf{Ir}_i(\mathbf{A}) \supset \mathbf{B}_i(A_i) \wedge \mathbf{B}_i(\mathbf{Ir}_j(\mathbf{A}))$.■

Because neither of IRA_i and IRI_i has the explicit inclusion of $\mathbf{Ir}_j(\cdot, \cdot)$, the operators $\mathbf{Ir}_i(\cdot, \cdot)$ and $\mathbf{Ir}_j(\cdot, \cdot)$ appear to be independent of one another. However, Lemma 2.2 shows that their

[9]Since the deduction theorem (cf., Mendelson [16]) does not hold in epistemic logic, the introduction of non-logical axioms differs from in classical logic. Our way is still following the logic literature. This is found in a comparion with a Gentzen-style formulation of epistemic logic (cf., Kaneko-Nagashima [11]).

interdependences are hidden in those axioms and rules. Actually, IRI_i allows to include $\mathbf{Ir}_j(\cdot,\cdot)$ in D_i. Now, we define the *epistemic content* of $\mathbf{Ir}_i(\mathbf{A})$ as follows;

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

The next lemma shows various properties of the operators $\mathbf{Ir}_i(\cdot;\cdot)$, $i = 1, 2$. We skip the proofs, which are given in Hu-Kaneko [7], Lemma 2.2. The properties (1), (2), and (4) are inherited from the belief operator $\mathbf{B}_i(\cdot)$ satisfying Axioms K and D and Rule NEC. In particular, (1) corresponds to NEC, (3) corresponds to Lemma 2.1.(5), and (4) corresponds to Axiom K. (2) reflects the self-referential structure. (5) follows from the negation of the Axiom IRA_i.

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* $\vdash C_k$ *for* $k = 1, 2$, *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) $\vdash \mathbf{Ir}_i(\neg A_i; A_j) \supset \neg\mathbf{Ir}_i(\mathbf{A})$, $\vdash \mathbf{Ir}_i(A_i; \neg A_j) \supset \neg\mathbf{Ir}_i(\mathbf{A})$, *and* $\vdash \mathbf{Ir}_i(\neg A_i; \neg A_j) \supset \neg\mathbf{Ir}_i(\mathbf{A})$.

The following statements for $\mathbf{Ir}_i^o(\cdot;\cdot)$ correspond to IRA_i and IRI_i for $\mathbf{Ir}_i(\cdot;\cdot)$, which play crucial roles in Sections 4 and 5.

Lemma 2.4. ***(IRA**$_i^o$ *and* **IRI**$_i^o$**)*** *Let* $\mathbf{A} = (A_i; A_j)$ *and* D_i *any formulae. Then,*

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

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

(2): Let $\vdash D_i \supset A_i \wedge \mathbf{B}_j(A_j) \wedge \mathbf{B}_j\mathbf{B}_i(D_i)$. 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 $\vdash D_i \supset \mathbf{Ir}_i^o(A_i; A_j)$ by (6).■

Although EIR^2 is the main logical system here, we mention some variants from time to time. Our undecidability result holds in stronger systems than EIR^2, such as those obtained from EIR^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/or Axiom 5 (negative introspection): $\neg\mathbf{B}_i(A) \supset \mathbf{B}_i(\neg\mathbf{B}_i(A))$[10]. In particular, when we add Axiom T to EIR^2, which is denoted by $\mathrm{EIR}^2(\mathrm{T})$, an infinite regress collapses to common knowledge. Lemma 2.2 implies $\vdash \mathbf{Ir}_i(A_1, A_2) \equiv \mathbf{Ir}_j(A_1, A_2)$ $(\equiv \mathbf{Ir}_i^o(A_1, A_2))$ for $i = 1, 2$ in $\mathrm{EIR}^2(\mathrm{T})$. It then holds in $\mathrm{EIR}^2(\mathrm{T})$ that for any formulae A_1, A_2 and D,

CKA: $\vdash \mathbf{Ir}_i(A_1, A_2) \supset (A_1 \wedge A_2) \wedge \mathbf{B}_1\mathbf{Ir}_i(A_1, A_2) \wedge \mathbf{B}_2\mathbf{Ir}_i(A_1, A_2)$;

CKI: if $\vdash D \supset (A_1 \wedge A_2) \wedge \mathbf{B}_1(D) \wedge \mathbf{B}_2(D)$, then $\vdash D \supset \mathbf{Ir}_i(A_1, A_2)$.

[10] We regard KD^2 as the basic system; Axiom K and Necessitation give the inference ability of classical logic to each player. Axiom D is needed for meaningful statements. If Axiom D is dropped, player's beliefs can be arbitrary with no restrictions. For example, it is proved with Axiom D but not wihout it that $\mathbf{B}_i(p) \nvdash \neg\mathbf{B}_i(\neg p)$.

Common Knowledge Logic can be formulated using the axiom and the inference rule corresponding to CKA and CKL (cf., Fagin et al. [4]). Since in EIR^2(T), CKA and CKI are derived formulae and admissible rule for $\mathbf{Ir}_i(A_1, A_2)$, $\mathbf{Ir}_i(A_1, A_2)$ means the common knowledge of $A_1 \wedge A_2$. However, we do not impose it unless stated otherwise, since Axiom T destroys players' subjective perspectives. See Hu-Kaneko [7] for a more detailed discussion.

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 either $i = 1, 2$. The set of nonepistemic formulae is denoted by $\mathcal{P}_N$. We will use the *belief eraser* ε_0 to connect EIR^2 to classical logic within $\mathcal{P}_N$. The nonepistemic formula $\varepsilon_0(A) \in \mathcal{P}_N$ is obtained from $A \in \mathcal{P}$ by eliminating all occurrences of $\mathbf{B}_1(\cdot), \mathbf{B}_2(\cdot)$ in A and replacing all occurrences of $\mathbf{Ir}_i(C_1, C_2)$ in A by $\varepsilon_0(C_1) \wedge \varepsilon_0(C_2)$ (formally, by induction). Then, we have

$$\vdash A \text{ implies } \vdash_0 \varepsilon_0(A), \tag{7}$$

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 its leaves (cf., Kaneko-Nagashima [10]).

2.3 Kripke semantics and the soundness/completeness of EIR^2

Here, we report soundness/completeness for EIR^2 with respect to the Kripke semantics. We will use the soundness part for the 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 $\langle (w_0, i_0), ..., (w_\nu, i_\nu), w_{\nu+1}\rangle \in (W \times \{1, 2\})^{\nu+1} \times W$ $(\nu \geq 0)$ is an *alternating chain* 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). We use these chains to evaluate the truth values of formulae $\mathbf{Ir}_i(A_1, A_2)$.

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 chain $\langle (w_0, i_0), ..., (w_\nu, i_\nu), w_{\nu+1}\rangle$ with $(w_0, i_0) = (w, i)$.

The above steps other than V6 are standard. V6 is similar to the valuation for the common knowledge operator in CKL; the difference is to use alternating reachability for two formulae, instead of simple rearchability (cf., Fagin *et al.* [4], 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 EIR^2 if and only if $(M, w) \models A$ for all serial models $M = (\langle W; R_1, R_2\rangle, \tau)$ and any $w \in W$.*

A proof of completeness is given in Hu-Kaneko [7] for the n-person case. In this paper, we use only soundness (only-if) for the undecidability result. Soundness is obtained 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 $\models$ preserves IRI_i. Both steps follow from V6.

Theorem 2.1 shows that our infinite regress operator $\mathbf{Ir}_i(\mathbf{A})$ faithfully captures the set in (5). The alternating rearchability 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. Moreover, 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 proof of Theorem 2.1 in [7] gives the (strong) finite model property (cf., p.145, 339, Blackburn, *et al.* [2]). Thus, EIR^2 is effectively decidable (simply "decidable" in the logic literature), i.e., the set of provable formulae is recursive. In Section 6, we discuss this problem relative to the game theoretic decidability/undecidability result for prediction/decision making.

The following lemma is crucial for interpretations of our decidability or undecidability results. In our applications, $\mathbf{B}_i(\Gamma_i^o)$ in the lemma takes the form $\mathbf{Ir}_i(\mathbf{C})$, and decidability takes the form $\mathbf{Ir}_i(\mathbf{C}) \vdash \mathbf{B}_i(A)$ or $\mathbf{Ir}_i(\mathbf{C}) \vdash \mathbf{B}_i(\neg A)$, as in (3). The lemma implies that this is equivalent to $\mathbf{Ir}_i(\mathbf{C}) \vdash \mathbf{B}_i(A)$ or $\mathbf{Ir}_i(\mathbf{C}) \vdash \neg\mathbf{B}_i(A)$.

Lemma 2.5. ***(Change of Scopes)****: Let Γ_i^o be a set of formulae and let A be a formula.*

(1) $\mathbf{B}_i(\Gamma_i^o) \vdash \mathbf{B}_i(A) \Longleftrightarrow \Gamma_i^o \vdash A$;

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

The direction $\Longleftarrow$ of (1) is immediate, and using Theorem 2.1, the contrapostive of $\Longrightarrow$ is proved by constructing a countermodel for $\mathbf{B}_i(\Gamma_i^o) \vdash \mathbf{B}_i(A)$. A detailed proof is given in Hu-Kaneko [7], Section 3.3. The direction $\Longrightarrow$ of (2) is proved as follows: The left-hand side is equivalent to $\mathbf{B}_i(\Gamma_i^o), \mathbf{B}_i(A) \vdash \mathbf{B}_i(\neg C \wedge C)$. By (1), $\Gamma_i^o, A \vdash \neg C \wedge C$, which is equivalent to $\Gamma_i^o \vdash \neg A$; so we have the right-hand side of (2). The converse is similar.

Lemma 2.5 requires KD^2 to be the base logic for EIR^2. If we add any of Axioms T, 4 or 5 to EIR^2, the lemma fails. Counterexamples are given in [7]; the failure means inseparability between player i's mind and the objective situation, which is incompatible with our basic motivation for our study of a player's individual independent decision making.

3 Game Theoretic Concepts and Some Completeness Results

Here, we give a few basic concepts in game theory relevant for our discussions, and formulate them in the language of EIR^2. We also prepare some completeness results for game formulae, which will be useful to understand our game theoretic undecidability result.

3.1 Basic game theoretic concepts

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 also 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 the set of all Nash equilibria in G by $E(G)$. The set $E(G)$ may be empty, e.g., Table 1.3 has the empty $E(G)$. We say that s_i is a *Nash strategy* iff $(s_i; s_j)$ is a Nash equilibrium for some $s_j \in S_j$.

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

$$\text{for all } s, s' \in E,\ (s_i; s'_j) \in E \text{ for } i = 1, 2. \tag{8}$$

This is equivalent to $E = E_1 \times E_2$, where $E_i = \{s_i \in S_i : (s_i; s_j) \in E \text{ for some } s_j\}$ for $i = 1, 2$. Let $\mathbf{E} = \{E : E \subseteq E(G) \text{ and } E \text{ satisfies (8)}\}$. The game G is *solvable* iff $E(G)$ satisfies (8), and then we call $E(G)$ the *Nash solution*. Otherwise, it is *unsolvable,* and 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 $F \subsetneq E'$. 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 is solvable but has the empty $E(G)$[11]. A sufficient condition for a game G to be solvable is that the payoffs are constant-sum, i.e., for some constant c, $h_1(s) + h_2(s) = c$ for all $s \in S$.

Hu-Kaneko [6] provided the axioms Na_1 and Na_2, which are given in Section 1, to characterize the Nash theory in the non-formalized language. The characterization of the Nash theory, given in [6], would be informative for our study in the logic EIR^2.

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$_2$*[12].

***(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 cases correspond basically to the game theoretic decidability and undecidability results to be given in the subsequent sections. Here, we avoided unnecessary complication for the case of $E(G) = \emptyset$. In the subsequent sections, however, we allow $E(G) = \emptyset$.

3.2 Some completeness results for game formulae

Since the players and strategies are already included in our formal language, it suffices to formalize payoff functions h_1 and h_2. They are expressed in terms of preference formulae as follows:

$$g_i = \wedge \left[\{\mathrm{Pr}_i(s; t) : h_i(s) \geq h_i(t)\} \cup \{\neg \mathrm{Pr}_i(s; t) : h_i(s) < h_i(t)\}\right]. \tag{9}$$

We call g_i the *formalized payoffs* associated with h_i for $i = 1, 2$. Since (9) also contains negative preferences, 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., under g_i, completeness holds for all atomic preference formulae for player i. The statement "$s_i \in S_i$ is a best response to $s_j \in S_j$" is expressed 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 $\mathrm{nash}(s) := \mathrm{bst}_1(s_1; s_2) \wedge \mathrm{bst}_2(s_2; s_1)$. We say that A is a *game formula for player i* iff the atomic formulae occurring in A are of the form $\mathrm{Pr}_i(\cdot; \cdot)$. The formula

[11] Nash [17] himself assumed the mixed strategies, and proved the existence of a Nash equilibrium. Here, we do not allow mixed strategies, and some games have no Nash equilibria.

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

$\mathrm{bst}_i(s_i; s_j)$ is a game formula for i, but $\mathrm{nash}(s)$ is not since it contains the atomic formulae of the form $\mathrm{Pr}_j(\cdot;\cdot)$.

Consistency of $g_1 \wedge g_2$ can be shown by constructing a truth assignment. Consistency of the infinite regress $\mathbf{Ir}_i(g_1, g_2)$ in EIR^2 is also 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, we have $g_1 \wedge g_2 \vdash_0 \neg\varepsilon_0 A \wedge \varepsilon_0 A$ by (7), which is impossible because of consistency of $g_1 \wedge g_2$. In the same way, we have consistency of $\mathbf{Ir}_i^o(g_1, g_2)$ in EIR^2. These are listed for the purpose of reference:

$$\mathbf{Ir}_i(g_1, g_2) \text{ and } \mathbf{Ir}_i^o(g_1, g_2) \text{ are consistent in } \mathrm{EIR}^2. \tag{10}$$

In addition to consistency, $\mathbf{Ir}_i(g_1, g_2)$ and $\mathbf{Ir}_i^o(g_1, g_2)$ enjoy certain completeness properties. They ensure that our game theoretic undecidability result obtained in Section 5 is not caused by a lack of content in the assumption set, and allow us to identify the crucial elements in the formalized Nash theory which cause undecidability.

As far as game formulae for players 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. *For $i = 1, 2$, let A_i be a nonepistemic game formula for player i. Let G be a game and $\mathbf{g} = (g_1, g_2)$ its formalized payoffs. Then,*

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

(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 $g_i \wedge \mathbf{Ir}_j(\mathbf{g}) \vdash A_i \wedge \mathbf{Ir}_j(\mathbf{A})$, i.e., $\mathbf{Ir}_i^o(\mathbf{g}) \vdash \mathbf{Ir}_i^o(\mathbf{A})$.

($(b) \Longrightarrow (c)$): We show its 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), $\mathbf{Ir}_i^o(\mathbf{g}) \vdash \mathbf{B}_j(\neg A_j)$; so $\mathbf{Ir}_i^o(\mathbf{g}) \vdash \neg\mathbf{B}_j(A_j)$ by Axiom D. But by Lemma 2.4.(1), $\vdash \mathbf{Ir}_i^o(\mathbf{A}) \supset \mathbf{B}_j(A_j)$, equivalently, $\vdash \neg\mathbf{B}_j(A_j) \supset \neg\mathbf{Ir}_i^o(\mathbf{A})$. Thus, $\mathbf{Ir}_i^o(\mathbf{g}) \vdash \neg\mathbf{Ir}_i^o(A_i; A_j)$. By (10), we have $\mathbf{Ir}_i^o(\mathbf{g}) \nvdash \mathbf{Ir}_i^o(A_i; A_j)$. The other cases are similar.■

The next theorem shows that $\mathbf{Ir}_i(\mathbf{g})$ is complete relative to infinite regresses of nonepistemic game formulae $\mathbf{A} = (A_1, A_2)$ for the players. We write the theorem in terms of the epistemic content $\mathbf{Ir}_i^o(\cdot;\cdot)$ for coherency of the later purpose.

Theorem 3.1. ***(Completeness for infinite regresses of game formulae)*** *For $i = 1, 2$, let A_i be a nonepistemic game formula for player i. Let G be a game and $\mathbf{g} = (g_1, g_2)$ its formalized payoffs. 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, $\mathbf{Ir}_i^o(\mathbf{g}) \vdash \neg\mathbf{Ir}_i^o(A_i; A_i)$.■

The completeness properties in Theorem 3.1 respect players' subjectivities; the epistemic layers of the assumption, $\mathbf{Ir}_i^o(\mathbf{g})$ coincide with those of the conclusion, $\mathbf{Ir}_i^o(\mathbf{A})$. This restriction is necessary because the logic EIR^2 maintains independence of players' minds. For example, $\mathbf{Ir}_i^o(\mathbf{g}) \nvdash \mathbf{B}_i(g_i)$ holds for EIR^2, which is proved by the epistemic separation theorem given Hu-Kaneko [7].

When we add Axiom T to EIR^2, we can remove the restriction; the result becomes the full completeness up to all game formulae, where we call A simply a *game formula* iff any atomic formula contained in A is of the form $\mathrm{Pr}_1(\cdot;\cdot)$ or $\mathrm{Pr}_2(\cdot;\cdot)$. The formula $\mathrm{nash}(s_1, s_2)$ is a nonepistemic game formula. The next theorem will be used for some results in Section 4.3.

Theorem 3.2. ***(Completeness for game formulae under Axiom T)*** *Let G be a game and $\mathbf{g} = (g_1, g_2)$ its formalized payoffs. For any (nonepistemic and epistemic) game formula A, either $\mathbf{Ir}_i(\mathbf{g}) \vdash A$ or $\mathbf{Ir}_i(\mathbf{g}) \vdash \neg A$ in EIR$^2(T)$.*

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

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, $\mathbf{B}_k(\mathbf{Ir}_i^o(\mathbf{g})) \vdash \mathbf{B}_k(C)$. If $k = j$, then $\mathbf{Ir}_i^o(\mathbf{g}) \vdash \mathbf{B}_j(\mathbf{Ir}_i^o(\mathbf{g}))$ by IRA_i^o and Axiom T; so $\mathbf{Ir}_i^o(\mathbf{g}) \vdash \mathbf{B}_k(C)$. If $k = i$, we have $\mathbf{Ir}_i^o(\mathbf{g}) \vdash \mathbf{B}_i(\mathbf{Ir}_i^o(\mathbf{g}))$ by IRA_i^o and Axiom T. In either case, we have, $\mathbf{Ir}_i^o(\mathbf{g}) \vdash \mathbf{B}_k(C)$. 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 $\mathbf{Ir}_i^o(\mathbf{g}) \vdash C_1 \wedge C_2$. By (6) and Axiom T, $\mathbf{Ir}_i^o(\mathbf{g}) \vdash \mathbf{Ir}_j^o(\mathbf{g})$ 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 $\mathbf{Ir}_i^o(\mathbf{g}) \vdash \mathbf{Ir}_k(C_1, C_2)$. When $\mathbf{Ir}_i^o(\mathbf{g}) \vdash (\neg C_i) \wedge C_j$, by the same argument, we have $\mathbf{Ir}_i^o(\mathbf{g}) \vdash \mathbf{Ir}_i(\neg C_i; C_j)$; so 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 Formalized Nash Theory

We give three axioms for player i's prediction/decision making, which formalize the decision criterion Na_i given in Section 3 taking beliefs into account. Then, we 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. In this section, we show, for a solvable 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. They are intended to be the contents of player i's basic beliefs and hence they occur in the scope of $\mathbf{B}_i(\cdot)$;

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

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

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

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 directly 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); this is explicitly formulated by $\mathrm{N1}_i$. The third states that in the mind of player i, his decision is correctly predicted by player j. This was only interpretational in Proposition 3.1. 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 $\mathbf{B}_j$.

Axioms N_i and N_j are interdependent: Since 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, $\mathbf{B}_i(\mathrm{I}_i(s_i))$ requires $\mathbf{B}_i\mathbf{B}_j\mathbf{B}_i(\mathrm{N}_i)$, and so on. To complete the prediction process the infinite regress formula $\mathbf{Ir}_i(\mathbf{N}) = \mathbf{Ir}_i(\mathrm{N}_i;\mathrm{N}_j)$ is needed. The infinite regress $\mathbf{Ir}_i(\mathbf{N})$ in the logic EIR^2 may be compared with Johansen's [9] interpretation of Nash theory, which will be discussed briefly in Section 6.

We take the infinite regress $\mathbf{Ir}_i(\mathrm{N}_i;\mathrm{N}_j)$ as basic beliefs for player i's prediction/decision making; $\mathrm{I}_i(s_i)$ and $\mathbf{B}_j(\mathrm{I}_j(s_j))$ in $\mathbf{Ir}_i(\mathrm{N}_i;\mathrm{N}_j)$ are treated as "unknowns" to be found by player i with his logical analysis. From $\mathbf{Ir}_i(\mathrm{N}_i;\mathrm{N}_j)$, necessary conditions for $\mathrm{I}_i(s_i)$ and $\mathrm{I}_j(s_j)$ are derived as the following game formulae: for each $i = 1, 2$ and $s_i \in S_i$,

$$A_i^*(s_i) := \vee_{t_j\in S_j}\mathbf{Ir}_i^o[\mathrm{bst}_i(s_i;t_j);\mathrm{bst}_j(t_j;s_i)]. \tag{11}$$

These candidate formulae play a crucial role in our subsequent analysis.

The nonepistemic content of $A_i^*(s_i)$ 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". In the logic $\mathrm{EIR}^2(\mathrm{T})$, we can interpret $\mathbf{Ir}_i(\cdot,\cdot)$ as the common knowledge operator, as stated in Section 2.2, and hence $A_i^*(s_i)$ means "s_i is a common knowledge Nash strategy". We emphasize this interpretation with Axiom T by writing $\mathbf{Ir}_i^o[\mathrm{bst}_i(s_i;t_j);\mathrm{bst}_j(t_j;s_i)]$ as $\mathbf{C}^*(\mathrm{Nash}(s_i;t_j))$ in $\mathrm{EIR}^2(\mathrm{T})$, and $A^*(s_i)$ is written as $\vee_{t_j\in S_j}\mathbf{C}^*(\mathrm{Nash}(s_i;t_j))$. This formula describes reality as well as both players' thinking, and it was adopted in Kaneko-Nagashima [10] and Kaneko [12]. Without Axiom T, the formula $A_i^*(s_i)$ occurs in the mind of player i, independent of reality as well as the other player j.

The following theorem gives a necessary condition for player i's possible final decisions, which will be proved in the end of this subsection. This "intermediate" step is useful to obtain negative decisions for both games with and without decidability results.

Theorem 4.1. ***(Necessity)*** *For $i = 1, 2$, $\mathbf{Ir}_i(\mathbf{N}) \vdash \mathbf{B}_i(I_i(s_i) \supset A_i^*(s_i))$ for all $s_i \in S_i$.*

That is, player i infers $A_i^*(s_i)$ as a necessary condition for his decision. By this 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 his 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. For our purposes, however, we only focus on implications of the form in Theorem 4.1.

Recalling $\varepsilon_0(A_i^*(s_i))$, the above theorem may be interpreted as meaning that a Nash strategy is derived. However, our target is prediction/decision making by a player. A possible decision resulting from this process is expressed by $\mathrm{I}_i(s_i)$, and $A_i^*(s_i)$ is only a necessary condition for it. This is a purely solution-theoretic statement in the sense that it does not depend upon payoffs. Also, even if payoffs, e.g., $\mathbf{Ir}_i(g_1, g_2)$, are specified, Theorem 4.1 does not give a positive answer to $\mathrm{I}_i(s_i)$; that is, its contrapositive may give only a negative decision $\neg \mathrm{I}_i(s_i)$ from $\neg A_i^*(s_i)$. We discuss the converse under the assumption of $\mathbf{Ir}_i(g_1, g_2)$ in later sections.

Here, we prove Theorem 4.1. It follows from (2) of the next lemma. (1) does not need $\mathrm{N1}_i$. We write $\mathrm{N0}_i \wedge \mathrm{N2}_i$ as $\mathrm{N02}_i$ 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): $\mathbf{Ir}_i^o[N_i; N_j] \vdash I_i(s_i) \supset A^*(s_i)$.

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 $i = 1, 2$,

$$\vdash \theta_i(s_i; s_j) \supset \mathrm{bst}_i(s_i; s_j) \wedge \mathbf{B}_j(\mathrm{bst}_j(s_j; s_i)) \wedge \mathbf{B}_j\mathbf{B}_i(\theta_i(s_i, s_j)). \tag{12}$$

By Lemma 2.4.(2), $\vdash \theta_i(s_i; s_j) \supset \mathbf{Ir}_i^o[\mathrm{bst}_i(s_i; s_j); \mathrm{bst}_j(s_i; s_j)]$; so we have the assertion.

The first part, $\vdash \theta_i(s_i; s_j) \supset \mathrm{bst}_i(s_i; s_j)$, of (12) 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 $\vdash \theta_i(s_i, s_j) \wedge \mathbf{B}_j(\mathrm{I}_j(s_j)) \wedge \mathbf{B}_j\mathbf{B}_i(\mathrm{I}_i(s_i)) \supset \mathbf{B}_j(\mathrm{bst}_j(s_j; s_i))$. Observe that $\mathbf{B}_j(\mathrm{I}_j(s_j))$ is included in $\theta_i(s_i, s_j)$ and $\mathbf{B}_j\mathbf{B}_i(\mathrm{I}_i(s_i))$ 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 (12). 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 $\vdash \mathbf{Ir}_i^o[\mathrm{N02}_i; \mathrm{N02}_j] \wedge \mathrm{I}_i(s_i) \supset \mathbf{B}_j\mathbf{B}_i(\mathrm{I}_i(s_i))$, and, using $\mathbf{B}_j(\mathrm{N2}_j)$ in $\mathbf{Ir}_i^o[\mathrm{N02}_i; \mathrm{N02}_j]$, we have $\vdash \mathbf{Ir}_i^o[\mathrm{N02}_i; \mathrm{N02}_j] \wedge \mathbf{B}_j(\mathrm{I}_j(s_j)) \supset \mathbf{B}_j\mathbf{B}_i\mathbf{B}_j(\mathrm{I}_j(s_j))$. Summing those three up, we obtain $\vdash \theta_i(s_i; s_j) \supset \mathbf{B}_j\mathbf{B}_i(\theta_i(s_i; s_j))$.

(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., $\mathbf{Ir}_i^o[\mathrm{N}_i; \mathrm{N}_j] \vdash \mathrm{I}_i(s_i) \supset A_i^*(s_i)$.■

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

The basic belief $\mathbf{Ir}_i[\mathrm{N}_i; \mathrm{N}_j]$ only gives necessary conditions for $\mathrm{I}_i(s_i)$ and $\mathbf{B}_j(\mathrm{I}_j(s_j))$, but not sufficient conditions. In fact, there are formulae, other than $A_i^*(s_i)$ and $A_j^*(s_j)$, enjoying the properties described by N_i and N_j. For example, the families of formulae, $\{\bot(s_i)\}_{s_i \in S_i}$, $i = 1, 2$, where $\bot(s_i) := \neg(p \supset p)$, $s_i \in S_i$ and p is an atomic preference formula, make $\mathrm{N}_i = \mathrm{N0}_i \wedge \mathrm{N1}_i \wedge \mathrm{N2}_i$ trivially hold with the substitution of $\bot(s_i)$ for each $\mathrm{I}_i(s_i)$ in N_i. To avoid such unintended candidates and to analyze the exact logical contents of $\mathbf{Ir}_i[\mathrm{N}_i; \mathrm{N}_j]$, we choose families of formulae $\{A_i(s_i)\}_{s_i \in S_i}$ and $\{A_i(s_j)\}_{s_j \in S_j}$ having the *exact* properties N_i and N_j.

We formalize this choice by an axiom scheme. We call $\mathcal{A} = (\mathcal{A}_i; \mathcal{A}_j)$ a pair of *candidate families* iff $\mathcal{A}_i = \{A_i(s_i)\}_{s_i \in S_i}$ and $\mathcal{A}_j = \{A_i(s_j)\}_{s_j \in S_j}$ are families of formulae 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 replacing all occurrences of $\mathrm{I}_k(s_k)$ in N_i by $A_k(s_k)$ for each $s_k \in S_k, k = 1, 2$. We denote the following formula by $\mathrm{WF}_i(\mathcal{A})$:

$$\begin{aligned} \mathrm{N}_i(\mathcal{A}) \wedge \mathbf{B}_j(\mathrm{N}_j(\mathcal{A})) \wedge [\wedge_{s\in S}\langle \mathrm{I}_i(s_i) \wedge \mathbf{B}_j(\mathrm{I}_j(s_j)) \quad &\supset \quad A_i(s_i) \wedge \mathbf{B}_j(A_j(s_j))\rangle] \\ &\supset \quad \wedge_{s_i\in S_i}\langle A_i(s_i) \supset \mathrm{I}_i(s_i)\rangle. \end{aligned} \tag{13}$$

Let $\mathbf{WF}(\mathcal{A}) = (\mathrm{WF}_1(\mathcal{A}),\mathrm{WF}_2(\mathcal{A}))$. The axiom scheme for the choice of the *deductively weakest formulae* is defined by $\mathbf{Ir}_i(\mathbf{WF}) := \{\mathbf{Ir}_i(\mathbf{WF}(\mathcal{A})) : \mathcal{A}$ is a pair of candidate families$\}$.

The formula in (13) contains the additional premise $\wedge_{s\in S}\langle \mathrm{I}_i(s_i) \wedge \mathbf{B}_j(\mathrm{I}_j(s_j)) \supset A_i(s_i) \wedge \mathbf{B}_j(A_j(s_j))\rangle$. 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. It is used together with $\mathbf{Ir}_i(\mathrm{N}_i;\mathrm{N}_j)$. Then the premise corresponds to the maximality requirement in the definition of a subsolution in Section 3. If we drop the premise, (13) becomes

$$\mathrm{WF}_i^+(\mathcal{A}) := \mathrm{N}_i(\mathcal{A}) \wedge \mathbf{B}_j(\mathrm{N}_j(\mathcal{A})) \supset \wedge_{s_i\in S_i}\langle A_i(s_i) \supset \mathrm{I}_i(s_i)\rangle. \tag{14}$$

This is stronger than $\mathrm{WF}_i(\mathcal{A})$. As we show later, it works only for a solvable game, but not for an unsolvable game, while $\mathrm{WF}_i(\mathcal{A})$ in (13) 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 $\mathbf{Ir}_i(\mathbf{g}) = \mathbf{Ir}_i(g_i; g_j)$. We postulate the entire set of axioms, denoted by $\Delta_i(\mathbf{g}) := \{\mathbf{Ir}_i(\mathbf{g}), \mathbf{Ir}_i(\mathbf{N})\}\cup \mathbf{Ir}_i(\mathbf{WF})$, as the basic beliefs for player i's prediction/decision making. The *formalized Nash theory* is expressed as $(\mathrm{EIR}^2; \Delta_i(\mathbf{g}))$. That is, we fix the logic EIR^2, and within it, we have the set of nonlogical axioms $\Delta_i(\mathbf{g})$. We are interested in the logical implications related to prediction/decision making derived from $\Delta_i(\mathbf{g})$ in EIR^2.

We state the consistency of the formalized Nash theory $(\mathrm{EIR}^2; \Delta_i(\mathbf{g}))$, which will be proved in the proof of Lemma 5.1.

Lemma 4.2. ***(Consistency of the belief set)*** *$\Delta_i(\mathbf{g})$ is consistent for any game G.*

If we replace $\mathbf{Ir}_i(\mathbf{WF})$ by $\mathbf{Ir}_i(\mathbf{WF}^+)$ given in (14), then $\Delta_i^+(\mathbf{g}) = \{\mathbf{Ir}_i(\mathbf{g}), \mathbf{Ir}_i(\mathbf{N})\}\cup\mathbf{Ir}_i(\mathbf{WF}^+)$ is consistent if and only if G is a solvable game, and Δ_i^+ is equivalent to Δ_i for any solvable G.

4.3 Game theoretic decidability for solvable games

Here, we show that the basic beliefs $\Delta_i(\mathbf{g})$ determine possible final decisions for a solvable game.

Theorem 4.2. ***(Determination I)*** *Let G be a solvable game and $\mathbf{g}$ its formalized payoffs. Then, for $i = 1, 2$, $\Delta_i(\mathbf{g}) \vdash \mathbf{B}_i(I_i(s_i) \equiv A_i^*(s_i))$ for all $s_i \in S_i$.*

Proof. We prove the following three claims.

Claim 1: Let G be solvable. Then, $\mathbf{Ir}_i^o(\mathbf{g}) \vdash A_i^*(s_i) \wedge \mathbf{B}_j(A_j^*(s_j)) \supset \mathrm{bst}_i(s_i; s_j)$.

Claim 2: $\vdash A_i^*(s_i) \supset \vee_{t_j\in S_j}\mathbf{B}_j(A_j^*(t_j))$. Claim 3: $\vdash A_i^*(s_i) \supset \mathbf{B}_j\mathbf{B}_i(A_i^*(s_i))$.

Proof of Claim 1: Since $\mathrm{bst}_i(s_i; s_j)$ is a game formula for $i = 1, 2$, we have, for each $s \in S$, $\mathbf{Ir}_i^o(\mathbf{g}) \vdash \mathbf{Ir}_i^o(\mathrm{bst}_i(s_i; s_j);\mathrm{bst}_j(s_j; s_i))$ or $\mathbf{Ir}_i^o(\mathbf{g}) \vdash \neg\mathbf{Ir}_i^o(\mathrm{bst}_i(s_i; s_j);\mathrm{bst}_j(s_j; s_i))$ by Theorem 3.1. Hence, for each $s_i \in S_i$, $\mathbf{Ir}_i^o(\mathbf{g}) \vdash A_i^*(s_i)$ or $\mathbf{Ir}_i^o(\mathbf{g}) \vdash \neg A_i^*(s_i)$. Using Lemma 2.2, we have, for each $s_j \in S_j$, $\mathbf{Ir}_i^o(\mathbf{g}) \vdash \mathbf{B}_j(A_j^*(s_j))$ or $\mathbf{Ir}_i^o(\mathbf{g}) \vdash \neg\mathbf{B}_j(A_j^*(s_j))$. Also, for each $s \in S$, $\mathbf{Ir}_i^o(\mathbf{g}) \vdash \mathrm{bst}_i(s_i; s_j)$ or $\mathbf{Ir}_i^o(\mathbf{g}) \vdash \neg\mathrm{bst}_i(s_i; s_j)$. Thus, $\mathbf{Ir}_i^o(\mathbf{g}) \vdash A_i^*(s_i) \wedge \mathbf{B}_j(A_j^*(s_j)) \supset \mathrm{bst}_i(s_i; s_j)$ or $\mathbf{Ir}_i^o(\mathbf{g}) \vdash$

$\neg[A_i^*(s_i) \wedge \mathbf{B}_j(A_j^*(s_j)) \supset \mathrm{bst}_i(s_i; s_j)]$. 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 a solvable game. Hence, we have the assertion.

Proof of Claim 2: 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., $\vdash A_i^*(s_i) \supset \vee_{t_j \in S_j} \mathbf{B}_j(A_j^*(t_j))$.

Proof of Claim 3: 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., $\mathbf{B}_j\mathbf{B}_i(A_i^*(s_i))$.

Here, we prove the theorem. It follows from the above claims that $\mathbf{Ir}_i^o(\mathbf{g}) \vdash \mathrm{N}_i(\mathcal{A}^*)$ for $i = 1, 2$. Hence, $\mathbf{Ir}_i^o(\mathbf{g}) \vdash \mathrm{N}_i(\mathcal{A}^*) \wedge \mathbf{B}_j(\mathrm{N}_j(\mathcal{A}^*))$. It follows from Theorem 4.1 that $\mathbf{Ir}_i^o(\mathrm{N}_i; \mathrm{N}_j) \vdash \wedge_{s \in S}[\mathrm{I}_i(s_i) \wedge \mathbf{B}_j(\mathrm{I}_j(s_j)) \supset A_i^*(s_i) \wedge \mathbf{B}_j(A_j^*(s_j))]$. Thus, $\mathbf{Ir}_i^o(\mathbf{g}), \mathbf{Ir}_i^o(\mathbf{N}), \mathbf{Ir}_i^o(\mathbf{WF}) \vdash A_i^*(s_i) \supset \mathrm{I}_i(s_i)$. Hence, $\Delta_i(\mathbf{g}) \vdash \mathbf{B}_i(A_i^*(s_i) \supset \mathrm{I}_i(s_i))$. Combining this with Theorem 4.1, we have the assertion of the theorem.■

Theorem 4.2 implies that $\Delta_i(\mathbf{g}) \vdash \mathbf{B}_i\mathbf{B}_j(\mathrm{I}_j(s_j) \equiv A_j^*(s_j))$ for all $s_j \in S_j$. That is, for a solvable game G, player i infers from his beliefs $\Delta_i(\mathbf{g})$ that his possible decision and prediction are fully expressed by $A_i^*(s_i)$ and $\mathbf{B}_j(A_j^*(s_j))$. As remarked above, in the logic $\mathrm{EIR}^2(\mathrm{T})$, $A_i^*(s_i)$ can be written as $\vee_{t_j \in S_j} \mathbf{C}^*(\mathrm{Nash}(s_i; t_j))$, and Theorem 4.2 becomes $\Delta_i(\mathbf{g}) \vdash \mathrm{I}_i(s_i) \equiv \vee_{t_j \in S_j} \mathbf{C}^*(\mathrm{Nash}(s_i; t_j))$. That is, a possible decision s_i is a Nash strategy with common knowledge. This corresponds to the result given in Kaneko [13], which assumes Axiom T, but here we extend the analysis to a purely subjective framework.

Then, because of the above theorem and Theorem 3.1, player i can decide whether a given strategy s_i is a final decision for him or not, which is stated by the following theorem.

Theorem 4.3. ***(Game theoretic decidability)*** *Let G be a solvable game and $\mathbf{g} = (g_1, g_2)$ its formalized payoffs. Then, for $i = 1, 2$ and each $s_i \in S_i$, either $\Delta_i(\mathbf{g}) \vdash \mathbf{B}_i(I_i(s_i))$ or $\Delta_i(\mathbf{g}) \vdash \mathbf{B}_i(\neg I_i(s_i))$.*

Proof. Since $\mathrm{bst}_i(s_i; s_j)$ is a nonepistemic game formula for $i = 1, 2$, it follows from Theorem 3.1 that $\mathbf{Ir}_i^o(\mathbf{g}) \vdash \mathbf{Ir}_i^o[\mathrm{bst}_i(s_i; s_j); \mathrm{bst}_j(s_j; s_i)]$ 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 $\mathbf{Ir}_i^o(\mathbf{g}) \vdash \mathbf{Ir}_i^o[\mathrm{bst}_i(s_i; s_j); \mathrm{bst}_j(s_j; s_i)]$ for some $s_j \in S_j$; so, $\mathbf{Ir}_i^o(\mathbf{g}) \vdash \vee_{t_j} \mathbf{Ir}_i^o[\mathrm{bst}_i(s_i; t_j); \mathrm{bst}_j(t_j; s_i)]$, i.e., $\mathbf{Ir}_i^o(\mathbf{g}) \vdash A_i^*(s_i)$. If not, we have $\mathbf{Ir}_i^o(\mathbf{g}) \vdash \neg \vee_{t_j} \mathbf{Ir}_i^o[\mathrm{bst}_i(s_i; t_j); \mathrm{bst}_j(t_j; s_i)]$, 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 Theorem 4.2, we have $\Delta_i(\mathbf{g}) \vdash \mathbf{B}_i(\mathrm{I}_i(s_i))$ or $\Delta_i(\mathbf{g}) \vdash \mathbf{B}_i(\neg \mathrm{I}_i(s_i))$.■

It holds for a solvable game G that for each strategy $s_j \in S_j$,

$$\text{either } \Delta_i(\mathbf{g}) \vdash \mathbf{B}_i\mathbf{B}_j(\mathrm{I}_j(s_j)) \text{ or } \Delta_i(\mathbf{g}) \vdash \mathbf{B}_i\mathbf{B}_j(\neg \mathrm{I}_j(s_j)). \tag{15}$$

Thus, player i can predict whether a given strategy s_j is a possible decision for player j for not. From now on, we concentrate on decidability or undecidability for player i.

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 Theorem 4.3 corresponds to whether s_i is a Nash strategy or not. For the negative recommendation, we need to

add only $\mathbf{Ir}_i(\mathbf{g})$ to $\mathbf{Ir}_i(\mathbf{N})$ in Theorem 4.1; that is, if s_i is not a Nash strategy, then

$$\mathbf{Ir}_i(\mathbf{g}), \mathbf{Ir}_i(\mathbf{N}) \vdash \mathbf{B}_i(\neg \mathrm{I}_i(s_i)). \tag{16}$$

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

Since Table 1.1 is a solvable game, Theorem 4.3 is applicable, and the belief set $\Delta_1(\mathbf{g})$ recommends strategy $\mathbf{s}_{12}$ as a positive decision to player 1, but $\mathbf{s}_{11}$, $\mathbf{s}_{13}$ as negative decisions. Table 1.2 is an unsolvable game; Theorem 4.2 is not applicable. In Table 1.3, (16) recommends all strategies as negative decisions.

Theorem 4.3 is enough for our purpose from the game theoretic perspective. At expense of the subjective nature for decision/prediction making, however, we obtain full completeness with Axiom T. As a corollary, $(\mathrm{EIR}^2(\mathrm{T}); \Delta_i(g))$ is effectively decidable.

Theorem 4.4. ***(Full Completeness with Axiom T)*** *Let G be a solvable game. Then, the theory $(EIR^2(T); \Delta_i(\mathbf{g}))$ is complete, i.e., for any $A \in \mathcal{P}$, $\Delta_i(\mathbf{g}) \vdash A$ or $\Delta_i(\mathbf{g}) \vdash \neg A$.*

Proof. In $\mathrm{EIR}^2(\mathrm{T})$, it holds that $\Delta_i(\mathbf{g}) \vdash \mathrm{I}_i(s_i) \equiv A_i^*(s_i)$ for any $s_i \in S_i$ and $i = 1, 2$. Let C be any formula, and $C^\#$ the formula obtained by replacing each occurrence of $\mathrm{I}_i(s_i)$ in C by $A_i^*(s_i)$ $(s_i \in S_i, i = 1, 2)$. We can show by induction of the length of a formula that $\Delta_i(\mathbf{g}) \vdash C^\# \equiv C$. We consider only the step of $C = \mathbf{Ir}_i(C_1, C_2)$. The induction hypothesis is that $\Delta_i(\mathbf{g}) \vdash C_k^\# \equiv C_k$ for $k = 1, 2$. Using IRA_i and T, $\Delta_i(\mathbf{g}) \vdash A$ implies $\Delta_i(\mathbf{g}) \vdash \mathbf{B}_k(A)$ for $k = 1, 2$ in $\mathrm{EIR}^2(\mathrm{T})$. It follows from IRA_i that $\Delta_i(\mathbf{g}) \vdash \mathbf{Ir}_i(C_1, C_2) \supset \mathbf{B}_i(C_i^\#) \wedge \mathbf{B}_i\mathbf{B}_j(C_j^\#) \wedge \mathbf{B}_i\mathbf{B}_j(\mathbf{Ir}_i(C_1, C_2))$. By IRI_i, we have $\Delta_i(\mathbf{g}) \vdash \mathbf{Ir}_i(C_1, C_2) \supset \mathbf{Ir}_i(C_1^\#, C_2^\#)$. The converse is parallel.

Then, since $\Delta_i(\mathbf{g}) \vdash C^\# \equiv C$ for any formula C and since $\Delta_i(\mathbf{g}) \vdash C^\#$ or $\Delta_i(\mathbf{g}) \vdash \neg C^\#$ by Theorem 3.2, we have $\Delta_i(\mathbf{g}) \vdash C$ or $\Delta_i(\mathbf{g}) \vdash \neg C$.■

5 Game Theoretic Undecidability for Unsolvable Games

In contrast to game theoretic decidability for solvable games, when G is unsolvable, we have the undecidability result that each player i has some strategy s_i so that he cannot infer from his belief set $\mathbf{\Delta}_i(\mathbf{g}) = \{\mathbf{Ir}_i(\mathbf{g}), \mathbf{Ir}_i(\mathbf{N})\} \cup \mathbf{Ir}_i(\mathbf{WF})$ whether s_i is a final decision or not. We are also able to obtain a full characterization of strategies satisfying game theoretic undecidability. In contrast to the decidable case where $\mathrm{I}_i(s_i)$ can be substantiated by a game formula (Theorem 4.2), for the undecidable case we show that there exist no such fomrulae.

5.1 Game theoretic undecidability

First we present our undecidability result. We give the proofs in Section 5.2.

Theorem 5.1. ***(Game theoretic undecidability)*** *Let G be an unsolvable game, and $\mathbf{g} = (g_1, g_2)$ its formalized payoffs. For $i = 1, 2$, there is an $s_i \in S_i$ such that*

$$\textit{neither } \mathbf{\Delta}_i(\mathbf{g}) \vdash \mathbf{B}_i(I_i(s_i)) \quad \textit{nor} \quad \mathbf{\Delta}_i(\mathbf{g}) \vdash \mathbf{B}_i(\neg I_i(s_i)). \tag{17}$$

By Lemma 2.5.(2), (17) is equivalent to $\mathbf{\Delta}_i(\mathbf{g}) \nvdash \mathbf{B}_i(\mathrm{I}_i(s_i))$ and $\Delta_i(g) \nvdash \neg\mathbf{B}_i(\mathrm{I}_i(s_i))$. Hence, $(\mathrm{EIR}^2; \Delta_i(\mathbf{g}))$ is incomplete. Theorem 5.1 also holds in $\mathrm{EIR}^2(\mathrm{T})$. Since $\vdash \mathbf{Ir}_i(A_1, A_2) \equiv \mathbf{Ir}_j(A_1, A_2)$ $(\equiv \mathbf{Ir}_i^o(A_1, A_2))$ in $\mathrm{EIR}^2(\mathrm{T})$, it implies that $\mathbf{\Delta}_i(\mathbf{g}) \nvdash \mathbf{B}_i(\mathrm{I}_i(s_i))$ and $\Delta_i(\mathbf{g}) \nvdash \neg\mathbf{B}_i(\mathrm{I}_i(s_i))$. Thus, $(\mathrm{EIR}^2(\mathrm{T}); \Delta_i(\mathbf{g}))$ is incomplete. This is in contrast to the solvable case, for which the theory is complete, as stated in Theorem 4.4.

Recall that Theorem 3.2 states, even for an unsolvable game, that $\mathbf{Ir}_i(\mathbf{g})$ is complete within the set of game formulae in $\mathrm{EIR}^2(\mathrm{T})$. Thus, when G is unsolvable, no game formulae can be used to express $\mathrm{I}_i(s_i)$ in the theory $(\mathrm{EIR}^2(\mathrm{T}); \mathbf{\Delta}_i(\mathbf{g}))$. This observation leads to the following theorem.

Theorem 5.2. ***(No-formula)*** *Let G be an unsolvable 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 (17) holds. Then, in $EIR^2(T)$, (also in EIR^2), there is no game formula A_i such that $\mathbf{\Delta}_i(\mathbf{g}) \vdash \mathbf{B}_i(I_i(s_i) \equiv A_i)$.*

In our language, game formulae describe the physical situation, while $\mathrm{I}_i(s_i)$ is introduced purely for decision making purposes wihout substantially meaning by itself. However, Theorem 5.2 shows that $\mathrm{I}_i(s_i)$ cannot be substantiated by any game formulae, and hence, to resolve the undecidability, interactions between the physical world and the decison-making process are necessary. One potential interaction is *ex post* observation, and we discuss this issue in Hu-Kaneko [7].

Returing to undecidability *per se*, Theorem 5.1 states existence of strategies satisfying (17), and here we give a full characterization of such strategies. The negative decision given in (16) holds for all non-Nash strategies s_i for any game G. Hence, s_i for (17) has to be a Nash strategy. A necessary and sufficient condition for (17) is that

$$s_i \text{ is a Nash strategy but } s_i \notin F_i \text{ for some subsolution } F_1 \times F_2. \tag{18}$$

The proof is given in the end of Section 5.2. In the battle of the sexes (Table 1.2), since this holds for both $\mathbf{s}_{i1}$ and $\mathbf{s}_{i2}$, $i = 1, 2$, we have undecidability (17) for both strategies of both players. This observation can be generalized as follows: when each subsolution is a singleton set, every Nash strategy s_i satisfies (18), and hence (17) holds for that strategy. A sufficient condition for each subsolution to be singleton is that all payoffs are distinct.

Without this condition, however, some Nash strategies may fail to satisfy (18). 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, (18) does not hold for $\mathbf{s}_{i1}$, but it holds for $\mathbf{s}_{i2}$. Our theorem below shows that for $i = 1, 2$, (17) holds for $\mathbf{s}_{i2}$ but not for $\mathbf{s}_{i1}$.

Table 5.1

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

Let G be any game with its subsolutions $F^1, ..., F^k$. We denote $\cap_{l=1}^k F^l$ by $\hat{F}$. We stipulate that if G has no Nash equilibria, then $k = 0$ and $\hat{F} = \emptyset$. If $k = 1$, then F^1 is the set of all Nash equilibria $E(G)$. This intersection $\hat{F}$ satisfies interchangeability (8); so it is written as $\hat{F} = \hat{F}_1 \times \hat{F}_2$. As stated above, when all payoffs are distinct, $\hat{F} = \cap_{l=1}^k F^l = \emptyset$ for $k \geq 2$.

We have the following characterization of the case of having a positive decision.

Theorem 5.3. ***(Positive Decision)*** *Let G be any game, $\mathbf{g} = (g_1, g_2)$ its formalized payoffs, and $i = 1, 2$. Then, for all $s_i \in S_i$, $\Delta_i(\mathbf{g}) \vdash \mathbf{B}_i(I_i(s_i))$ if and only if $s_i \in \hat{F}_i$.*

This theorem has various implications. First, when G has no Nash equilibria, i.e., $\hat{F} = \emptyset$, $\Delta_i(\mathbf{g})$ gives negative decisions for all strategies. When G is solvable, it gives a positive decision to each Nash strategy and a negative decision to any other strategy. When G has multiple subsolutions, there are two cases: if $\hat{F} = \emptyset$, then it gives no positive decision; if $\hat{F} \neq \emptyset$, it gives a positive decision for each $s_i \in \hat{F}_i$.

5.2 Proof of the theorems

We stipulate that when $E(G) = \emptyset$, then the subsolution F is empty and $F_1 = F_2 = \emptyset$. The proof of Lemma 5.1 together with soundness for EIR^2 also gives a proof of Lemma 4.2.

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*

$$(M, w) \models \mathbf{Ir}_i(\mathbf{g}) \wedge \mathbf{Ir}_i(\mathbf{N}) \text{ and } (M, w) \models \mathbf{Ir}_i(\mathbf{WF}(\mathcal{A})) \text{ for all } \mathcal{A}; \tag{19}$$

$$\text{for any } s_i \in S_i,\ (M, w) \models \mathbf{B}_i(I_i(s_i)) \Leftrightarrow (M, w) \models I_i(s_i) \Leftrightarrow s_i \in F_i. \tag{20}$$

Proof. We construct a model $M = (\langle W; R_1, R_2\rangle, \tau)$ satisfying (19) and (20). 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. Define τ by, for $k = 1, 2$,

$$\text{for any } s; s' \in S,\ \tau(\mathrm{PR}_k(s; s')) = \top \Leftrightarrow h_k(s) \geq h_k(s'); \tag{21}$$

$$\tau(w, \mathrm{I}_k(s_k)) = \top \Leftrightarrow s_k \in F_k. \tag{22}$$

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 (21), we have $(M, w) \models g_1 \wedge g_2$. Also, since $W = \{w\}$, we have, for any formula C and $k = 1, 2$,

$$(M, w) \models C \ \Leftrightarrow (M, w) \models \mathbf{B}_k(C). \tag{23}$$

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 (22), 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 (23), that if $(M, w) \models \neg[\mathrm{N}_1(\mathcal{A}) \wedge \mathrm{N}_2(\mathcal{A})]$, then $(M, w) \models \mathrm{WF}_i(\mathcal{A})$. 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) Let $E_1 \times E_2 \subseteq F$. Then, by (22), for $k = 1, 2$, $(M, w) \models \wedge_{s_k \in S_k}[A_k(s_k) \supset \mathrm{I}_k(s_k)]$; so $(M, w) \models \mathrm{WF}_i(\mathcal{A})$.

(ii) Let $E_1 \times E_2 - F \neq \emptyset$. 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, $(M, w) \models \mathrm{WF}_i(\mathcal{A})$ for $i = 1, 2$.■

Proof of Theorem 5.1: Let G be an unsolvable game, and let F, F' be two subsolutions with $(s_i; s_j) \in F$ but $(s_i; s_j) \notin F'$ for $s = (s_i; s_j)$. By Lemma 5.1, there are two models M and M' so that (19) and (20), 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 EIR^2, we have $\Delta_i(\mathbf{g}) \not\vdash \neg\mathbf{B}_i(\mathrm{I}_i(s_i))$ and $\Delta_i(\mathbf{g}) \not\vdash \mathbf{B}_i(\mathrm{I}_i(s_i))$.■

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

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

The necessity in Theorem 5.3 requires a modification of the previous characterization (Theorem 4.2). 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. Thus, $A^{**}(s_i)$ depends upon a game G. We define the candidate formulae $\mathcal{C}_i = \{C_i^*(s_i)\}_{s_i \in S_i}, i = 1, 2$ as follows:

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

That is, $C_i^*(s_i)$ is $A_i^{**}(s_i)$ if $s_i \in \hat{F}_i$, but is $A_i^*(s_i)$ if s_i is not a Nash strategy. It is crucial to set $C_i^*(s_i)$ to be $\mathrm{I}_i(s_i)$ if s_i is a Nash strategy but is not a part of the intersection $\hat{F}$. The last treatment trivializes the additional premise in WF_i of (13). Then, we have an extension of Theorem 4.2.

Lemma 5.2. ***(Determination 2)*** *Let G be a game with its subsolutions $F^1, ..., F^k$, and $\mathbf{g} = (g_1, g_2)$ its formalized payoffs. For $i = 1, 2$, $\mathbf{\Delta}_i(\mathbf{g}) \vdash \mathbf{B}_i(I_i(s_i) \equiv C_i^*(s_i))$ for all $s_i \in S_i$.*

Proof. 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,

$$\mathbf{Ir}_i^o(\mathbf{g}), \mathbf{Ir}_i^o(\mathbf{N}) \vdash \ \mathrm{I}_i(s_i) \supset C_i^*(s_i) \text{ for all } s_i \in S_i. \tag{26}$$

Now, consider the converse of (26).

We modify the claims 1-3 in the proof of Theorem 4.2 as follows: for any $(s_i; s_j) \in S$,

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

(2*): $\mathbf{Ir}_i^o(\mathbf{g}), \mathbf{Ir}_i^o(\mathbf{N}) \vdash C_i^*(s_i) \supset \vee_{t_j \in S_j} \mathbf{B}_j(C_j^*(t_j))$. (3*): $\mathbf{Ir}_i^o(\mathbf{N}) \vdash C_i^*(s_i) \supset \mathbf{B}_j\mathbf{B}_i(C_i^*(s_i))$.

(1*): 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.

(2*): 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 (26) 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, $\vdash C_i^*(s_i) \supset \vee_{t_j \in S_j} \mathbf{B}_j(C_j^*(t_j))$.

(3*): 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 the previous claim 3. 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 $\mathrm{N2}_i$.■

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 (26), 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 5.3: (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, $\Delta_i(\mathbf{g}) \vdash \mathbf{B}_i(\neg \mathrm{I}_i(s_i))$ by (16); so $\Delta_i(\mathbf{g}) \vdash \neg \mathbf{B}_i(\mathrm{I}_i(s_i))$ by Axiom D. Since $\Delta_i(\mathbf{g})$ is consistent by Lemma 4.2, we have $\Delta_i(\mathbf{g}) \nvdash \mathbf{B}_i(\mathrm{I}_i(s_i))$. 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, $\Delta_i(\mathbf{g}) \nvdash \mathbf{B}_i(\mathrm{I}_i(s_i))$ by (18).

(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, $\Delta_i^o(\mathbf{g}) \vdash \mathrm{I}_i(s_i)$ by Theorem 5.2, which implies $\Delta_i(\mathbf{g}) \vdash \mathbf{B}_i(\mathrm{I}_i(s_i))$.■

Proof of (18): Necessity: If s_i is not a Nash strategy, we have $\Delta_i(\mathbf{g}) \vdash \mathbf{B}_i(\neg \mathrm{I}_i(s_i))$ by (16). If $s_i \in \hat{F}_i$, then $\Delta_i(\mathbf{g}) \vdash \mathbf{B}_i(\mathrm{I}_i(s_i))$ by Theorem 5.3. Sufficienty: Suppose that s_i is a Nash strategy but $s_i \notin F_i$ for some subsolution $F_1 \times F_2$. Again, by Theorem 5.3, $\Delta_i(\mathbf{g}) \nvdash \mathbf{B}_i(\mathrm{I}_i(s_i))$. By Lemma 5.1 and its (20), we have a reflexive model $M = (\langle W; R_1, R_2 \rangle, \tau)$ of $\Delta_i(\mathbf{g})$ such that $W = \{w\}$ and $(M, w) \nvDash \neg \mathrm{I}_i(s_i))$. By Theorem 4.4, we have $\Delta_i(\mathbf{g}) \nvdash \mathbf{B}_i(\neg \mathrm{I}_i(s_i))$.■

6 Conclusions

We have considered prediction/decision making by player i in a finite 2-person game G. We describe his decision criterion as $\mathrm{N}_i = \mathrm{N0}_i \wedge \mathrm{N1}_i \wedge \mathrm{N2}_i$, which occurs in his mind, with the symmetric treatment for player j. These lead to an infinite regress of N_i and N_j, formalized by $\mathbf{Ir}_i(\mathbf{N}) = \mathbf{Ir}_i(\mathrm{N}_i; \mathrm{N}_j)$ in EIR^2. We have adopted $\mathbf{Ir}_i(\mathbf{N})$ as his basic beliefs, together with $\mathbf{Ir}_i(\mathbf{WF})$ and $\mathbf{Ir}_i(\mathbf{g})$. For a solvable game G, $\Delta_i(\mathbf{g}) = \{\mathbf{Ir}_i(\mathbf{g}), \mathbf{Ir}_i(\mathbf{N})\} \cup \mathbf{Ir}_i(\mathbf{WF})$ determines $\mathrm{I}_i(s_i)$ as the specific formula $A^*(s_i)$ given in (11). The situation for an unsolvable G is entirely different: for some strategy s_i, $\Delta_i(\mathbf{g})$ fails to determine whether it is a possible final decision or not. Here, we discuss these game theoretic decidability and undecidability results, comparisons to the literature, and some possible extensions.

Positive, negative decisions, and undecidable: When G is solvable, our decidability states that player i finds any Nash strategy to be a possible decision, and any non-Nash strategy to be a negative decision. Depending on the game, player i may find multiple possible final decisions or no positive decision. In the former case, our theory is silent for further choice from multiple decisions, and in the latter, negative decisions due to emptiness may lead player i to a different decision criterion.

In contrast, when G is unsolvable, we presented the undecidability result that player i cannot find any positive decision, unless the subsolutions have the nonempty intersection. One potential

solution is to strengthen the prediction/decision criterion for player i or to allow communication between the players so that they may agree upon a specific subsolution. In the first place, however, player i may not notice the necessity of such possibilities.

Two independent minds and discord in $\mathbf{Ir}_i(\mathbf{g})$: Theorem 5.1 implies the incompleteness of the theory $(\mathrm{EIR}^2; \boldsymbol{\Delta}_i(\mathbf{g}))$ (and even $(\mathrm{EIR}^2(\mathrm{T}); \boldsymbol{\Delta}_i(\mathbf{g}))$, when G is solvable. This is parallel to Gödel's incompleteness theorem. These two incompleteness results have some similarity but their sources are different.

Gödel's theorem is caused by the self-referential structure of Peano Arithmetic, i.e., the theory of Peano Arithmetic can be described inside the theory itself. Our framework also includes a self-referential structure. The infinite regress operator $\mathbf{Ir}_i(\cdot;\cdot)$ includes $\mathbf{Ir}_j(\cdot;\cdot)$, and *vice versa* in EIR^2. Moreover, the criteria $\{\mathbf{Ir}_i(\mathbf{N})\} \cup \mathbf{Ir}_i(\mathbf{WF})$ are completely symmetric between the two minds. Our undecidability arises in this context, but it is not directly generated by this structure. The direct cause lies in the infinite regress of the game $\mathbf{Ir}_i(\mathbf{g})$, which includes a possible discord between the players, depending upon whether the game is solvable or not.

Johansen's argument: This situation may be better understood by looking at Johansen's [**9**] argument. He gave the following four postulates for prediction/decision making and asserted that the Nash noncooperative solution could be derived for solvable games.

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 strategy 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[13] 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.

These postulates, except for J2, correspond directly to $\mathrm{N0}_i$,$\mathrm{N1}_i$,$\mathrm{N2}_i$ for $i = 1, 2$. Postulate J2 is interpreted as corresponding to the self-referential structure described above. That is, player i assumes the entirely symmetric structure for player j's thinking; complete symmetry is obtained in terms of infinite regresses $\{\mathbf{Ir}_i(\mathbf{N})\} \cup \mathbf{Ir}_i(\mathbf{WF})$ in the logic EIR^2, while still keeping the independence of the two minds. Once $\mathbf{Ir}_i(\mathbf{g})$ is introduced, it may contain some discord. However, Johansen did not discuss this part.

Undecidable cases and *ex post* observations: In EIR^2, we can allow the two players to have totally different basic beliefs; they may have different prediction/decision criteria and/or different beliefs about the game being played. Indeed, Hu-Kaneko [7] considers this possibility explicitly and shows that in EIR^2 the two players' minds can be fully separated. However, the prior basic beliefs may be upset by *ex post* observations, and the interaction between *ex ante* prediction/decision making and *ex post* observations may lead to new frontiers for game theory. See also Hu-Kaneko [7] for some discussions about belief-consistency and *ex post* observations.

Other game theoretic undecidability: Kaneko-Nagashima [10] consider a 3-person game with integer payoffs having a unique Nash equilibrium in mixed strategies. It is assumed that

[13] This "any" was "some" in Johansen's original Postulate 3. He assumed (p.435) that the game has the unique Nash equilibrium. In this case, the above difference does not matter.

the game structure and real number theory Φ_{rcf} (real closed field theory) are common knowledge among the players in an infinitary predicate logic. They show that the players can commonly know the abstract existence of a Nash equilibrium, but do not find a concrete one; hence they cannot play the specific Nash equilibrium strategy. This undecidability is caused by the lack of names for some irrational numbers such as $\sqrt{51}$ in their language, which is involved in the Nash equilibrium in the 3-person game. In contrast to our undecidability result, the main difficulty there is to give a name to a concept (i.e., $x^2 = 51$ and $x \geq 0$), but not the self-referential structure.

Other game theoretic solution concepts: Our undecidability result requires the infinite regress of beliefs. In particular, it may not apply to other decision criteria based on various "solution concepts" (cf., Osborne-Rubinstein [18]) other than the Nash theory. One example is the "dominant strategy" criterion, which requires a player to choose a strategy that best responses against any strategy of the other player. We may also extend this criterion by requiring one player to use a best response against any dominant strategy of the other, predicting that he adopts the dominant strategy criterion. Even we can extend this argument to any finite depth, starting with the dominant strategy criterion at the deepest level. In those cases, we have game theoretic decidability result. We conjecture that any solution concept which does not require infinite regress will lead to similar decidability[14].

Effective decidability of the theory: When G is a solvable game, effective decidability of the theory $(\mathrm{EIR}^2(\mathrm{T}); \Delta_i(\mathbf{g}))$ follows from the full completeness theorem (Theorem 4.2). For $(\mathrm{EIR}^2; \Delta_i(\mathbf{g}))$, we need to restrict the class of formulae. When G is unsolvable, this argument does not work: the effective decidability in such a case remains open.

Future directions: Our approach assumes unbounded logical abilities and unbounded interpersonal thinking, but we still meet the undecidability result. From the social science perspective, it may be fruitful to investigate whether a theory with bounded logical abilities or bounded interpersonal thinking can avoid undecidability. This is an entirely open problem.

References

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

[2] Blackburn, P., M. de Rijke, and T. Venema, (2001), *Modal Logic*, Cambridge University Press, Cambridge.

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

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

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

[6] 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

[14] The concept called "rationalizability" (cf., Osborne-Rubinstein [18]) involves an infinite regress of believes. For this concept, we do not meet undecidability, since it does not allow the other to have independent thinking.

[7] Hu, T., and M. Kaneko (2014), Epistemic Infinite Regress Logic, to be completed in 2014.

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

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

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

[11] Kaneko, M., and T. Nagashima, (1997), Game logic and its applications II, *Studia Logica* 58, 273–303.

[12] Kaneko, M., (1999), Epistemic considerations of decision making in games. *Mathematical Social Sciences* 38, 105–137.

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

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

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

[20] Van Benthem, J. *Modal Logic for Open Minds*, CSLI Lecture Notes, Standford, CA: CSLI Publication,

[21] Van Benthem, J., *Logic in Games*, The MIT Press, Cambridge, Massachusetts.

[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.

[23] Venema, Y., (2007), Lectures on the modal μ-calculus, Institute for Logic, Language and Computation, University of Amsterdam.