Interpersonal Beliefs and Inferences: the Role of Truth Axiom and Logical Consistency

Tai-Wei Hu*

January 14, 2022

Abstract

A framework for epistemic analysis of solution concepts in games based on proof-theoretical analysis is proposed. Players form beliefs about one another's substantive rationality and infer each other's actions from such beliefs. In this system, this inference can be explicitly formalized. In particular, a belief system that characterizes level-k theory is formalized in this system and is shown to be logically consistent in a system where players have classical logical abilities without imposing the truth axiom. However, once the truth axiom is imposed, the level-k theory is consistent if and only if the default actions form a Nash equilibrium for all $k \geq 2$. Finally, with an additional stationary axiom for beliefs in substantive rationality, infinite regress of beliefs in substantive rationality is consistent with Nash equilibrium only, regardless of whether truth axiom is imposed or not.

Key words:

*University of Bristol, email: taiwei.hu@bristol.ac.uk

1 Introduction

The paper proposes a new framework to describe and analyse strategic thinking in games. Modern economic models have gradually moved from price-taking behaviour in general equilibrium models to strategic considerations in game theory. At the heart of strategic analysis is a model of interpersonal reasoning. However, in the standard formulation this model is informal and is coupled with solution concepts. A leading example is Nash equilibrium, the most popular solution concept in most applied works, with the implicit assumption that players can anticipate other players' equilibrium behaviour. In recent years, however, many researchers are more concerned with the decision-making process. A case in point is level-k theory, typically regarded as a behavioural approach in which a bound on interpersonal reasoning is imposed. However, most works focus on the *outcomes* of various behavioural assumptions on interpersonal reasoning instead of directly modelling the reasoning process.

I focus on the interpersonal reasoning processes in game situations and analyze different requirements and properties in terms the players' logical abilities and logical consistency for the two most notable solution concepts, level-k theory and Nash equilibrium. The framework is based on the epistemic logic developed in Hu et al. (2019), which takes a syntactical approach and is capable of explicitly describing each player's interpersonal beliefs and logical inferences. It differs from the standard models of epistemic game theory (as surveyed in Dekel and Siniscalchi, 2015) in that while standard type space formulation begins with arbitrary "epistemic types," or hierarchies of players' higher-order beliefs, my framework focuses on the *process* that leads to those beliefs, starting with a formal description of substantive rationality in games and interpersonal beliefs upon that rationality.

By modelling the inference process explicitly, I am able to disentangle some implicit assumptions/restrictions in the traditional models with implications to the resulting solution concepts. The basic framework only requires players to have standard logical abilities as described by the classical propositional logic and to be free from logical inconsistency in their beliefs. In particular, I do not impose the truth axiom, the positive introspection axiom, or the negative introspective axioms on the players' beliefs. This approach allows for a formal level-k theory as describing a situation where the player is

capable of making inferences based on interpersonal belief of substantive rationality up to level-k to make predictions about his/her opponent's action. At the deepest level, the level-k theory appeals to a default choice that is given, which functions as the *anchor* of interpersonal beliefs. I first show that taking these as basic beliefs, a player can infer the action as predicted by the level-k theory; moreover, all the basic beliefs are logically consistent and hence level-k theory has a coherent foundation based on substantive rationality.

However, it turns out that the lack of truth axiom is crucial for this coherence, unless the default actions are prescribed according to the Nash equilibrium. More formally, under the truth axiom, the basic beliefs for level-k theory are logically consistent if and only if the default actions constitute a Nash equilibrium. Intuitively, truth axiom requires beliefs to be verifiable w.r.t. the objective situation. In contrast, the level-k theory requires interpersonal beliefs at different layers to be independent of one another, since at each layer, the assumption about the bound on the depth of interpersonal reasoning differs. Thus, level-k theory is not coherent with substantive rationality and interpersonal beliefs about it only when coupled with the truth axiom. Moreover, these results also show that, when coupled with the truth axioms, level-k theory is coherent if and only if the default actions constitute a Nash equilibrium.

Here I illustrate these results in the context of the undercutting game in Costa-Gomes and Crawford (2006), a version of which is described in Table 1. According to the level-k theory with default actions (a_4, b_4), a level-1 row player would take action a_3 (undercutting), a level-2 column player would take b_2, and a level-3 row player would take a_1, as well as a level-4 column player taking b_1. Since (a_1, b_1) is a Nash equilibrium, all level-k players for $k \geq 3$ play that equilibrium. In the context of this game, the above results imply that level-k theory is inconsistent with the truth axiom when $k \geq 2$, even though the outcome coincides with equilibrium behaviour. This highlights the importance of the *process* in my framework, not just the outcome. Instead, the level-k theory would be consistent with the truth axiom is the default choices are (a_1, b_1) to begin with, i.e., equilibrium is prescribed to all levels.

Finally, I show that the coherence with the truth axiom can be restored if the bound on interpersonal beliefs in level-k theory is removed. More precisely, I replace the cognitive bound by an infinite regress of interpersonal belief on the substantive rationality,

	b_1	b_2	b_3	b_4
a_1	$(1,1)$	$(1,0)$	$(0,0)$	$(0,0)$
a_2	$(0,1)$	$(0,0)$	$(1,0)$	$(0,0)$
a_3	$(0,0)$	$(0,1)$	$(0,0)$	$(1,0)$
a_4	$(0,0)$	$(0,0)$	$(0,1)$	$(0,0)$

Table 1: The undercutting game

and replace the default choice by a stationarity requirement in the interpersonal beliefs. I then show that the Nash actions as the only course of actions logically consistent with the infinite regress of these basic beliefs. In coordination game, this result does not solve the coordination problem but predicts that logical consistency requires one of the Nash equilibrium to be played. In contrast, if the game satisfies the interchangeability assumption required by Nash (1950), then players can infer that the Nash actions as the definite outcomes from these beliefs. In generic games, interchangeability essentially requires the game to have a unique Nash equilibrium.

Related literature

The level-k theory is developed to describe findings from experiments on game situations (see, e.g., Stahl and Wilson, 1994, 1995, Nagel, 1995, Costa-Gomes et al., 2001, Camerer et al. 2004). The focus here is of a more theoretical nature, and the goal is to clarify coherence of the level-k theory with various axioms for rationality as defined in economics. However, my results may contribute to the experimental literature by pointing out the importance of the truth axiom and logical consistency in level-k reasoning. The lack of truth or the existence of self-contradictory beliefs are not uncommon in people's beliefs, but here I point out that these "behavioural" traits are necessary for the level-k theory to explain behaviour deviating from Nash.

The dominant approach to study epistemology in games follows the type structures by Harsanyi (1967), as surveyed by Dekel and Siniscalchi (2014). Following this approach, Kets (2014) proposes a type-structure to capture reasoning with finite depths. More recently, Brandenburger et al. (2020) constructs a level-k type structure to cap-

ture the level-k reasoning. The main conclusion from there, however, is that level-k reasoning is intrinsically different from the type-space approach to epistemology in games. In contrast, by adopting a proof-theoretical system, my approach can directly capture the level-k reasoning. Moreover, my approach directly pinpoints the main consistency issue in the level-k theory, namely, it cannot be logically consistent once the truth axiom is imposed. In this sense, my results also compliment those in Aumann and Brandenburger (1995), who require mutual knowledge of the actions being played and payoff-maximization behaviour for Nash equilibrium. Under the truth axiom, my result says that only Nash equilibrium survives logical consistency when the players have at least two layers of interpersonal beliefs of rationality, a result similar in flavour to Aumann and Brandenburger (1995). However, my result also discuss the case where the truth axiom is not imposed and how one can obtain Nash equilibrium without prior knowledge of the other player's action.

The epistemic logic employed here is originated from Fagin et al. (1995), who provide the completeness theorem for various systems including the finitary KD system used here. The proof-theoretic approach to study solution concepts in game theory starts with Kaneko and Nagashima (1996, 1997), but the focus has been on infinite reasoning depths and Nash solution. Feinberg (2005) also uses a proof-theoretical framework to analyze epistemic conditions in extensive-form games, where logical consistency is also on the forefront to understand the logic of backward induction. My results for unbounded depths of interpersonal beliefs are obtained using the small infinitary epistemic logic developed by Hu et al. (2020). Finally, some of the results in Section 5 regarding Nash equilibrium and unbounded depths of reasoning are directly borrowed from Hu and Kaneko (2014), but here I reinterpret those results in terms of logical consistency and contrast them with level-k theory from the perspective of the truth axiom.

2 The Model

To model beliefs and interpersonal inferences, I begin with an abstract framework before applying it to the game situation. The formalism offers a clear-cut distinction between the underlying environment and the agent's decision-process for a meaningful

analysis of interaction between the two in terms of outcomes.

2.1 The epistemic system KD

The system begins with a description of the underlying language, whose core element is a set of *atomic propositions*, denoted by $\mathcal{P}_0$. Each atomic proposition is a true/false description of the underlying environment in a given economic model. To describe beliefs and logical inferences, we form other propositions that start with $\mathcal{P}_0$ by using the following logical connectives and belief operators:

logical connective symbols: $\neg$ (not), $\Rightarrow$ (imply), $\wedge$ (and), $\vee$ (or);[1]

unary belief operators: $\mathbf{B}_i(\cdot)$, $i = 1, 2, ..., N$.

The sets of *propositions*, denoted by $\mathcal{P}$, is then obtained by induction:

(o) all atomic propositions are propositions;

(i) if $A,\ B$ are propositions, so are $(A \Rightarrow B)$, $(\neg A)$, $\mathbf{B}_i(A)$, $i = 1, ..., N$;

(ii) if Φ is a finite (nonempty) set of propositions, $(\wedge\Phi)$ and $(\vee\Phi)$ are propositions.[2]

I write $\wedge\{A, B\}$, $\wedge\{A, B, C\}$ as $A \wedge B$, $A \wedge B \wedge C$, etc., and $(A \Rightarrow B) \wedge (B \Rightarrow A)$ as $A \equiv B$. I abbreviate parentheses or use different ones such as $[,]$. I also write $\wedge\mathbf{B}_i(\Phi)$ for $\wedge\{\mathbf{B}_i(A) : A \in \Phi\}$, etc.

I impose typical axioms in classical propositional logic. These consist of five axiom (schemata) and three inference rules: for all propositions A, B, C, and finite nonempty sets Φ of propositions,

L1 $A \Rightarrow (B \Rightarrow A)$; **L2** $(A \Rightarrow (B \Rightarrow C)) \Rightarrow ((A \Rightarrow B) \Rightarrow (A \Rightarrow C))$;

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

L4 $\wedge\Phi \Rightarrow A$, where $A \in \Phi$; and **L5** $A \Rightarrow \vee\Phi$, where $A \in \Phi$;

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

These axioms and inference rules capture a rational agent's basic logical abilities, and I assume that the agents can use them to make their own logical inferences, that is, the agents possess *instrumental rationality*. To model this, however, note that the

[1] 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.

[2] I presume the identity of finite sets in our language.

axioms and rules above only describe the inferences from the perspective of an *outside observer*. To describe the instrumental rationality of the agents *in the model*, whose beliefs and inferences have to be described within the scope of the belief operator $\mathbf{B}_i$, axioms and rules w.r.t. those operators are needed: for all formulae A, C, and for $i = 1, 2$,

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

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

As will be seen later, these axioms and inference rules ensure that the agents enjoy logical abilities as described by classical propositional logics and suppose that other agents enjoy the same. Axiom K implies that if the implication described $A \Rightarrow B$ is believed by the agent, and if the agent believes in the premise A, then the agent must be about to infer proposition B and believes in it. Axiom D also ensures that no agent has contradictory beliefs. Finally, Necessitation states that if A is a logical tautology, then the agent can infer it himself/herself, and believes in it.[3]

A *proof* $\langle X, <; \psi\rangle$ 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. I call $\langle X, <; \psi\rangle$ *a proof of* A iff $\psi(x_0) = A$, where x_0 is the root of $\langle X, <\rangle$. I say that A is *provable*, denoted by $\vdash A$, iff there is a proof of A.

For later purposes, I also need to consider the corresponding semantics of KD. We use the standard Kripke model, denoted by $M = (W, R_1, ..., R_N, \tau)$, which consists of

- set of possible worlds, W;
- accessibility relation for each i, R_i,
- truth valuation, $\tau : W \times \mathcal{P}_0 \to \{\top, \bot\}$.

[3]Note that, however, I do not impose the truth axiom, nor any introspection (positive or negative).

The truth evaluation is then extended to all formulae $\neg A, A \Rightarrow B, A \wedge B, \mathbf{B}_i(A) \in \mathcal{P}$ as follows:

- $\tau(w, A) = \top$ iff $\tau(w, \neg A) = \bot$,
- $\tau(w, A \wedge B) = \top$ iff $\tau(w, A) = \top = \tau(w, B)$,
- $\tau(w, A \Rightarrow B) = \top$ iff $\tau(w, A) = \bot$ or $\tau(w, B) = \top$,
- $\tau(w, \mathbf{B}_i(A)) = \top$ iff $\tau(v, A) = \top$ for all v such that $(w, v) \in R_i$.

A formula A is *valid* (denoted $\models_M A$) under M iff $\tau(w, A) = \top$ for all w. We have the following Completeness Theorem (c.f. Fagin et al.) In a model M, the accessibility relation R_i is *serial* if for all $w \in W$, there exists v such that $(w, v) \in R_i$. The following result is well-known (see, e.g., Kaneko (2002)).

Theorem 2.1. *For any formulae $A \in \mathcal{P}$,*

$$\vdash A \text{ if and only if } \models_M A \tag{1}$$

for all M in which W_1, ..., W_N are serial.

2.2 The single-agent case

Here I consider a simple single-agent decision problem to illustrate the use of the framework, and to make two points. The first is to highlight the distinction between the agent i's mind and the outside observer, a distinction achieved by dedicating the agent's mental activities within the scope of the belief operator $\mathbf{B}_i$. The second is to model *substantive rationality*; here it means the principles that the agent use to make his/her decisions.

I consider a simple decision-making problem that requires some inferences as follows. The nature chooses some parameter $\theta \in \Theta$, a finite set, and then, upon observing the realization of θ, the agent chooses an action $a \in A$, also a finite set, with the utility function given by $u(a, \theta)$. To describe this environment in the above logical system, I use the following atomic propositions. For each $\theta \in \Theta$, I use the proposition $\mathrm{R}(\theta)$ to mean that θ realizes. For each pair of outcomes, (a, θ) and (a', θ'), I use $\Pr_i(a, \theta; a', \theta')$

to mean that the outcome (a, θ) is preferred to (a', θ'). Note that both $R(\theta)$ and $\Pr_i(a, \theta; a', \theta')$ are *propositions about the environment,* which the agent treat as datum for her decision. For each a, I use $\mathrm{D}_i(a)$ to mean that the agent i takes a as the intended action. The proposition $\mathrm{D}_i(a)$ is a *proposition about the agent's own intention,* which the agent has to decide/infer from her principles.

For a given utility function u, it can be described by the following proposition:

$$\Pr_u = (\wedge\{\Pr_i(a, \theta; a', \theta') : u(a, \theta) \geq u(a', \theta')\}) \wedge (\wedge\{\neg\Pr_i(a, \theta; a', \theta') : u(a, \theta) < u(a', \theta')\}).$$

Now we are ready to define substantive rationality. Here it means the following proposition:

$$\mathbf{SRN}_i = \bigwedge_{a \in A} \left\{ \mathrm{D}_i(a) \Rightarrow \bigwedge_{\theta \in \Theta} \{(\mathrm{R}(\theta) \Rightarrow \mathrm{BR}_i(a; \theta))\} \right\}, \tag{2}$$

where $\mathrm{BR}_i(a; \theta)$ abbreviates the following proposition:

$$\mathrm{BR}_i(a; \theta) = \wedge \{\Pr_i(a, \theta; a', \theta) : a' \in A\},$$

that is, action a is a best response when the realized state of nature is θ.

Thus, the proposition SRN states that an intended action must maximize the agent's payoff given the realized state of nature. The name SRN stands for "substantive rationality (necessity)": It only gives a necessary condition for an intended action, which is indicated by the connective $\Rightarrow$ right after $\mathrm{D}_i(a)$. As such, it can only be used to exclude certain actions, but not to positively affirm an action. The following proposition also requires a positive affirmation:

$$\mathbf{SR}_i = \bigwedge_{a \in A} \left\{ \mathrm{D}_i(a) \equiv \bigwedge_{\theta \in \Theta} \{(\mathrm{R}(\theta) \Rightarrow \mathrm{BR}_i(a; \theta))\} \right\}, \tag{3}$$

which states that, in addition to what is required in SRN, if an action a satisfies the condition

$$\hat{\mathrm{D}}_i(a) = \wedge_{\theta \in \Theta} \{(\mathrm{R}(\theta) \Rightarrow \mathrm{BR}_i(a; \theta))\},$$

then $\mathrm{D}_i(a)$ must hold, that is, a is an intended action. Note that the statement $\hat{\mathrm{D}}_i(a)$ only consists of propositions regarding the environment, without any proposition regarding the intention.

Up to now the propositions are expressed in objective terms. However, the proposition SR_i occurs in the agent's mind to consider his/her intended decision, as well as the realization $\mathrm{R}(\theta)$ and the payoff function Pr_u. For a given utility function u, let $a(\theta) = \arg\max_{a \in A} u(a, \theta)$. Here I assume a genericity condition so that the maximizer is unique for each θ. We have the following theorem:

$$\mathbf{B}_i(\mathrm{SR}_i \wedge \mathrm{R}(\theta) \wedge \mathrm{Pr}_u) \vdash \mathbf{B}_i(\mathrm{D}_i(a(\theta))) \wedge (\wedge\{\mathbf{B}_i(\neg \mathrm{D}_i(a)) : a \neq a(\theta)\}). \tag{4}$$

In economic theory, such derivation has two steps. The first one is an abstract *solution theory*: for all $a \in A$,

$$\mathbf{B}_i(\mathrm{SR}_i) \vdash \mathbf{B}_i(\mathrm{D}_i(a)) \equiv \mathbf{B}_i(\hat{\mathrm{D}}_i(a)). \tag{5}$$

That is, it reduces the requirement of the intended action to a property of that action according to the environment, say, the payoff function or the outcome determination. The second step then uses the belief about the environment to infer the actual intended actions:

$$\mathbf{B}_i(\mathrm{R}(\theta) \wedge \mathrm{Pr}_u) \vdash \mathbf{B}_i(\hat{\mathrm{D}}_i(a(\theta))) \wedge (\wedge\{\mathbf{B}_i(\neg \hat{\mathrm{D}}_i(a)) : a \neq a(\theta)\}). \tag{6}$$

Typically, to prove such theorems in economics, the outside observer or the analyst simulates the thinking or the inferences of the agent. The following result, proved in Hu et al. (2019), formally endorses this methodology.

Theorem 2.2. *For any proposition A and B,*

$$A \vdash B \text{ if and only if } \mathbf{B}_i(A) \vdash \mathbf{B}_i(B). \tag{7}$$

3 Substantive rationality in games

Here I use the framework to discuss two popular solution concepts, level-k theory and the Nash solution. For simplicity, I restrict my attention to a two-person normal-form game, $G = \langle (A_1, A_2), (u_1, u_2) \rangle$, where A_i is the set of actions for player i and $u_i : A_1 \times A_2 \to \mathbb{R}$ is the payoff function for player i, $i = 1, 2$. I only focus on games that satisfy the following genericity condition: for $i = 1, 2$, and for all a_j,

$$\text{there exists } a_i \text{ such that } u_i(a_i; a_j) > u_i(a_i'; a_j) \text{ for all } a_i' \neq a_i. \tag{8}$$

In other words, the best response is unique. Note that the game in Table 1 satisfies the assumption (8). To use the epistemic logic system KD in Section 2.1 with $N = 2$ to study interpersonal reasoning in the game G, I consider the following language based on the set of atomic propositions that consists of

preference propositions: $\mathrm{Pr}_i(a; a')$ for $i = 1, 2$ and $a, a' \in A$; and

decision propositions: $\mathrm{D}_i(a_i)$ for $a_i \in A_i$, $i = 1, 2$.

Note that these propositions are natural generalizations of those in Section 2.2 to the game situation. The atomic formula $\mathrm{Pr}_i(a; a')$ means that player i *weakly prefers* the outcome $a = (a_1, a_2)$ to the outcome $a' = (a'_1, a'_2)$. The atomic formula $\mathrm{D}_i(a_i)$ expresses the idea that, from player i's perspective, a_i is a *possible final decision* for him. As in Section 2.1, from these atomic propositions the set of all propositions is generated inductively. When all the atomic propositions in a proposition consist of preference propositions only, I call such a proposition a *game-proposition*. A game-proposition only includes description about the game situation, but does not involve any decision.

Given the game $G = \langle (A_1, A_2), (u_1, u_2) \rangle$, one can formalize the payoff functions u_1 and u_2 as follows:

$$g_i = \wedge \left(\{ \mathrm{Pr}_i(a; a') : u_i(a) \geq u_i(a') \} \cup \{ \neg \mathrm{Pr}_i(a; a') : u_i(a) < u_i(a') \} \right). \tag{9}$$

I call g_i the *formalized payoffs* associated with u_i, $i = 1, 2$. Clearly, g_i is a game-proposition. Since (9) also contains negative preferences, for all $a, a' \in A$,

$$g_i \vdash \mathrm{Pr}_i(a; a') \text{ or } g_i \vdash \neg \mathrm{Pr}_i(a; a'),$$

i.e., under g_i, completeness holds for all atomic preference proposition for player i. The statement "$a_i \in A_i$ is a best response against $a_j \in A_j$" is expressed as

$$\mathrm{Bst}_i(a_i; a_j) := \wedge_{a'_i \in A_i} \mathrm{Pr}_i((a_i; a_j); (a'_i; a_j)).$$

Note that $\mathrm{Bst}_i(a_i; a_j)$ is a game-proposition as well.

Now I turn to substantive rationality, which extends the propositions (2) and (3)

to accommodate the game situation:

$$\mathbf{GRN}_i = \bigwedge_{a_i \in A_i} \left(\mathrm{D}_i(a_i) \Rightarrow \left(\bigwedge_{a_j \in A_j} \left(\mathbf{B}_j(\mathrm{D}_j(a_j)) \Rightarrow \mathrm{Bst}_i(a_i; a_j) \right) \right) \right); \tag{10}$$

$$\mathbf{GR}_i = \bigwedge_{a_i \in A_i} \left(\mathrm{D}_i(a_i) \equiv \left(\bigwedge_{a_j \in A_j} \left(\mathbf{B}_j(\mathrm{D}_j(a_j)) \Rightarrow \mathrm{Bst}_i(a_i; a_j) \right) \right) \right). \tag{11}$$

Compared against $\mathbf{SRN}_i$ and $\mathbf{SR}_i$, the key difference in $\mathbf{GRN}_i$ and $\mathbf{GR}_i$ is that it substitutes $R(\theta)$ by $\mathbf{B}_j(\mathrm{D}_j(a_j))$, with the obvious reason that player j's action also affects i's payoff. The formulation where $\mathrm{D}_j(a_j)$ occurs within the scope of $\mathbf{B}_j$ reflects the fact that $\mathrm{D}_j(a_j)$ is in player j's control, and has to be determined by player j's substantive and instrumental rationality. In player i's mind, therefore, the determination of $\mathbf{B}_j(\mathrm{D}_j(a_j))$ has to come from his beliefs about player j's substantive and instrumental rationality.

In what follows I formulate such beliefs for the level-k theory and for the Nash solution. Both require interpersonal beliefs but they differ in the depths of such beliefs.

4 Level-k theory

The level-k theory begins with players of level-0. Level-0 players use some default actions. Specifically, I use $\bar{a}_i \in A_i$ to denote the default action used by a level-0 player i.[4] A level-1 player then believes that his opponent is a level-0 player and best responds to his belief. One can then inductively define a level-$(k+1)$ player as follows: a level-$(k+1)$ player believes that his opponent is a level-k player and best responds to his belief.

This theory delivers a solution concept parametrized by the deepest level, k, and the default actions, $(\bar{a}_1, \bar{a}_2)$, that is, for any given k and $(\bar{a}_1, \bar{a}_2)$, there is a behavioural prediction for a level-k player i. Under condition (8), this prediction is in fact uniquely determined as follows. Fix a player i and construct a sequence of actions, $\mathbf{a}_i = (a_i^1, a_j^2, ...)$,

[4]Since I consider only pure strategies, technically speaking mixed strategy is not allowed. However, I can introduce mixed strategies directly if we restrict players to use mixed strategies with probabilities on a finite grid.

$j \neq i$, as follows:

$$u_i(a_i^1, \bar{a}_j) \geq u_i(a_i, \bar{a}_j) \text{ for all } a_i \in A_i; \tag{12}$$

$$u_j(a_j^{k+1}, a_i^k) \geq u_j(a_j, a_i^k) \text{ for all } a_j \in A_j \text{ for each } k \geq 1 \text{ odd}; \tag{13}$$

$$u_i(a_i^{k+1}, a_j^k) \geq u_i(a_i, a_j^k) \text{ for all } a_i \in A_i \text{ for each } k \geq 2 \text{ even}. \tag{14}$$

Thus, (12) says that a_i^1 is the best response to $\bar{a}_j$, and by (8) it is uniquely determined. Similarly, (13) says that a_j^2 is the best response to a_i^1, and (14) says that a_i^3 is the best response to a_j^2, and so on. By (8) the whole sequence is uniquely determined. Then, the solution for a level-1 player i is a_i^1, for a level-2 player j is a_j^2, and so on. One can construct the corresponding sequence $\mathbf{a}_j = (a_j^1, a_i^2, ...)$ that begins with the best response a_j^1 to $\bar{a}_i$, and the solution for a level-1 player j is a_j^1, the solution for a level-2 player i is a_i^2, and so on, in a symmetric manner.

Informally, the level-k theory gives the following predictions. A level-k player i uses action a_i^k, with the belief that his opponent j is of level-$(k-1)$ and uses action a_j^{k-1}, a belief that is informally justified by inferring his opponent's action from his opponent's interpersonal reasoning as a level-$(k-1)$ player. To achieve that, player i needs to simulate player's j's inference, which in turn involves interpersonal reasoning of shallower depths. The action a_j^{k-1} appears in the next level, within the scope of player i's thinking about player j, where a_j^{k-1} is believed to be his opponent's intended action, a level-1 interpersonal belief reasoning. The process continues until player i reaches the kth level of interpersonal reasoning, which is his cognitive bound for such inferences, and he uses default choice $\bar{a}_{i_k}$ to determine the imaginary player's action at that level. Note that in this process, level-$(k-1)$ player occurs in the first layer of interpersonal reasoning, level-$(k-2)$ player in the second layer, and so on, until one reaches level-0 player (who uses default choice) at the deepest layer, k.

Now I give a preview of the formal epistemic analysis of the level-k theory. First I give some accounting of interpersonal inferences. For a level-k player i, the 0th layer of interpersonal reasoning occurs within the scope of $\mathbf{B}_i$, and it is within that level player i aims to determine his possible final decision, a_i^k. To determine a_i^k, player i uses his substantive rationality, $\mathbf{GR}_i$, together with his information about the payoffs, g_i. These basic beliefs are formulated as $\mathbf{B}_i(\mathbf{GR}_i \wedge g_i)$. However, as evident in (11), to determine a_i form $\mathbf{GR}_i$ a prediction $\mathbf{B}_j(\mathrm{D}_j(a_j))$ for player i is necessary, which would be expressed

as $\mathbf{B}_i\mathbf{B}_j(\mathrm{D}_j(a_j))$. To infer player j's possible final decisions, denoted by a_j^{k-1}, I assume that player i believes in player j's substantive rationality and information about the payoffs, and these beliefs can be expressed by $\mathbf{B}_i\mathbf{B}_j(\mathbf{GR}_j \wedge g_j)$. This process, however, cannot stop here. To determine a_j form $\mathbf{GR}_j$ a prediction $\mathbf{B}_i(\mathrm{D}_i(a_i))$ for player j is necessary, which then leads to basic beliefs $\mathbf{B}_i\mathbf{B}_j\mathbf{B}_i(\mathbf{GR}_i \wedge g_i)$.

In principle, the, to solve for a_i^k within the utmost layer, an infinite regress of interpersonal beliefs is necessary. Level-k theory, however, assumes that player i can only entertain interpersonal reasoning up to the kth layer. This also implies that at the kth layer of interpersonal belief the decision is not determined by the substantive rationality, but by default choices, formulated as: for player $i = 1, 2$,

$$\mathbf{DF}_i = \bigwedge\{\mathrm{D}_i(a_i) \equiv \mathrm{L}_i^0(a_i)\}, \tag{15}$$

where $\mathrm{L}_i^0(a_i) := A \vee (\neg A)$ if $a_i = \bar{a}_i$ and $\mathrm{L}_i^0(a_i) := A \wedge (\neg A)$ otherwise.

In summary, if player i is of level-k, his basic interpersonal beliefs regarding substantive rationality from layers 0 to $k-1$ include

$$\mathbf{B}_{i_0}(\mathbf{GR}_{i_0} \wedge g_{i_0}) \wedge \mathbf{B}_{i_0}\mathbf{B}_{i_1}(\mathbf{GR}_{i_1} \wedge g_{i_1}) \wedge \cdots \wedge \mathbf{B}_{i_0}\mathbf{B}_{i_1}\cdots\mathbf{B}_{i_{k-1}}(\mathbf{GR}_{i_{k-1}} \wedge g_{i_{k-1}}), \tag{16}$$

where $i_0 = i$, $i_t \neq i_{t-1}$ for all $t = 1, ..., k-1$. (Hence, $i_{k-1} = i$ if k is odd and $i_{k-1} = j$ if k is even.) At the deepest layer, however, the basic belief is

$$\mathbf{B}_{i_0}\mathbf{B}_{i_1}\cdots\mathbf{B}_{i_k}(\mathbf{DF}_{i_k}). \tag{17}$$

To simplify notations, I will use the notation $e_\ell^i = (i_0, ..., i_\ell)$, where $i_0 = i$, $i_t \neq i_{t-1}$, $t = 1, ..., \ell$, and will denote

$$\mathbf{B}_{e_\ell^i} = \mathbf{B}_{i_0}\mathbf{B}_{i_1}\cdots\mathbf{B}_{i_\ell}.$$

This formulation highlights two aspects of bounded rationality in the level-k theory. The first aspect is the depths of interpersonal reasoning. If player i would be "perfectly" rational, he would be able to continue the process with the interpersonal beliefs

$$\bigwedge_{\ell=0}^{\infty} \mathbf{B}_{e_\ell^i}(\mathbf{GR}_{i_\ell} \wedge g_{i_\ell})$$

with each layer using the substantive rationality of the imaginary player. We will in fact encounter this infinite regress in the discussion of the Nash theory below. Instead,

the level-k theory cuts this infinite regress at the kth layer and insert the default choice criterion given by (15). This last point brings about the second aspect, the choice of the default actions. As will be seen later, without the truth axiom (as within the KD system), the default choices can be arbitrary without being inconsistent. In contrast, with truth axiom, the choice of default actions will be connected to Nash equilibrium, a connection that is robust to the choice of $k \geq 2$.

To describe the solutions to the inferences from the interpersonal beliefs (16) and (17), define a sequence of propositions,

$$\widehat{D}^k_{i_k}(a_{i_k}) = \mathrm{L}^0_{i_k}(a_{i_k}) \text{ for all } a_{i_k} \in A_{i_k}; \tag{18}$$

and for $\ell = k-1, ..., 0$,

$$\widehat{D}^\ell_{i_\ell}(a_{i_\ell}) = \bigwedge_{a_{i_{\ell+1}} \in A_{i_{\ell+1}}} \left(\mathbf{B}_{i_{\ell+1}}(\widehat{D}^{\ell+1}_{i_{\ell+1}}(a_{i_{\ell+1}})) \Rightarrow \mathrm{Bst}_{i_\ell}(a_{i_\ell}; a_{i_{\ell+1}}) \right) \text{ for all } a_{i_\ell} \in A_{i_\ell}. \tag{19}$$

Two remarks regarding the propositions defined in (19) are in order. First, in the proposition $\widehat{D}^k_{i_k}(a_{i_k})$, only preference propositions are used but not any decision proposition is used at all, that is, $\widehat{D}^k_{i_k}(a_{i_k})$ is a game-proposition. Second, the superscript ℓ is meant to match the occurrence of the proposition in the layer of interpersonal reasoning. Thus, as will be seen below, $\widehat{D}^k_{i_k}(a_{i_k})$ occurs in the deepest layer, k, within the scope of $\mathbf{B}_{e^i_k}$. Similarly, $\widehat{D}^\ell_{i_\ell}(a_{i_\ell})$ occurs within the scope of $\mathbf{B}_{e^i_\ell}$ for $\ell = k-1, ..., 0$. Note that the convention that $i_0 = i$ is maintained throughout. We have the following theorem.

Theorem 4.1 (Solution theoretic characterization of level-k theory)**.**

$$\left(\bigwedge_{\ell=0}^{k-1} \mathbf{B}_{e^i_\ell}(\mathbf{GR}_{i_\ell}) \right) \wedge \mathbf{B}_{e^i_k}(\mathbf{DF}_{i_k}) \vdash \mathbf{B}_i \left(\bigwedge_{a_i \in A_i} (D_i(a_i) \equiv \widehat{D}^0_i(a_i)) \right). \tag{20}$$

The result (20) is solution-theoretic in the sense that the characterization does not depend on the specific payoff function, but characterizes the decisions using a game-proposition. The proof (which can be found in the Appendix) relies on the following result, which is proved by induction in m.

$$\left(\bigwedge_{\ell=m}^{k-1} \mathbf{B}_{e^{i_m}_{\ell-m}}(\mathbf{GR}_{i_\ell}) \right) \wedge \mathbf{B}_{e^{i_m}_{k-m}}(\mathbf{DF}_{i_k}) \vdash \mathbf{B}_{i_m} \left(\bigwedge_{a_{i_m} \in A_{i_m}} (\mathrm{D}_{i_m}(a_{i_m}) \equiv \widehat{D}^m_{i_m}(a_{i_m})) \right). \tag{21}$$

My next result shows that the interpersonal beliefs (16) and (17) lead to the specific actions according to a_i^k given by (12)-(14) when the player incorporates the beliefs about the payoff function.

Theorem 4.2 (Recommendation by the level-k theory)**.** *Suppose that G satisfies (8) and let g_1 and g_2 be the corresponding proposition given by (9). Then, for any $k \geq 1$,*

$$\left(\bigwedge_{\ell=0}^{k-1} \mathbf{B}_{e_\ell^i}(\mathbf{GR}_{i_\ell} \wedge g_{i_\ell})\right) \wedge \mathbf{B}_{e_k^i}(\mathbf{DF}_{i_k}) \vdash \mathbf{B}_i(D_i(a_i^k)) \wedge \left(\bigwedge_{a_i \neq a_i^k} \equiv \mathbf{B}_i(\neg D_i(a_i))\right), \quad (22)$$

where a_i^k is determined by (12)-(14). Moreover, the assumptions in (22) is consistent.

Theorem 4.2 shows that the level-k theory is well supported by substantive rationality formulated by $\mathbf{GR}_1$ and $\mathbf{GR}_2$, and cutting the interpersonal reasoning at the deepest level by the default actions. Moreover, it shows that the assumptions, namely the k layers of interpersonal beliefs, are logically consistent. Hence, substantive rationality imposed so far cannot determine the default choices and they can be chosen arbitrarily. An important factor to this result is the subjective nature of the system KD. Indeed, although the players have perfect logical abilities and consistency, they are allowed to have their subjective views without interaction between one another in the sense of the objective environment. More precisely, the formulation allows them to interact with their imaginary opponents and the imaginary opponents in their minds alone.

To introduce interaction with the objective world, below I will impose the truth axiom, and show that once it is introduced, the default actions can be determined to be equilibrium actions by consistency requirements.

Consistency of Level-k theory under truth axiom

As mentioned earlier, up to now the default choices are arbitrary. Here I use logical consistency to consider their determination. In doing so the role of a commonly used axiom will be highlighted, the truth axiom: for all proposition A,

$$\mathbf{B}_i(A) \Rightarrow A.$$

When the truth axiom is incorporated, I use $\vdash_T$ to denote the inferences. The truth axiom imposes a connection between players' beliefs to the outside observer's, and requires the player's belief to be true from the outside observer's perspective. In an environment with more than one player like ours, it also requires consistency of beliefs across different players. We have the following theorem.

Theorem 4.3. *Assume the genericity condition and Axiom T.*

(1) For $k = 1$, $\mathbf{B}_i(\mathbf{GR}_i \wedge g_i) \wedge \mathbf{B}_i\mathbf{B}_j(\mathbf{DF}_j)$ *is consistent.*

(2) For $k > 1$, $\left(\bigwedge_{\ell=0}^{k-1} \mathbf{B}_{e_\ell^i}(\mathbf{GR}_{i_\ell} \wedge g_{i_\ell})\right) \wedge \mathbf{B}_{e_k^i}(\mathbf{DF}_{i_k})$ *is consistent if and only if* (a_i^k, a_j^{k-1}) *is a Nash equilibrium, with* $\bar{a}_i = a_i^k$ *for* k *odd and* $\bar{a}_j = a_j^{k-1}$ *for* k *even.*

According to Theorem 4.3, $\mathbf{B}_i(\mathbf{GR}_i \wedge g_i) \wedge \mathbf{B}_i\mathbf{B}_j(\mathbf{DF}_j)$ is consistent even under Axiom T. The proof in fact shows that the consistency result will hold even if we impose, in addition to Axiom T, Axioms 4 and 5, that is, the standard S5 system. As a result, one can construct a model for level-1 players in a standard partition model.

In contrast, when $k > 1$, $\left(\bigwedge_{\ell=0}^{k-1} \mathbf{B}_{e_\ell^i}(\mathbf{GR}_{i_\ell} \wedge g_{i_\ell})\right) \wedge \mathbf{B}_{e_k^i}(\mathbf{DF}_{i_k})$ is consistent under Axiom T (as the proof shows, under the standard S5 system as well) if and only if the final decision (a_i^k) and the prediction (a_j^{k-1}) forms a Nash equilibrium. Under the genericity condition, this can happen if and only if players of all levels play the same Nash equilibrium and the default choice is part of the equilibrium.[5] As a result, the cognitive bound has no bite on predictions and the outcome coincides with Nash equilibrium. In other words, the Nash outcome is robust to any bound $k \geq 2$. This also implies that, if the default choice is not a Nash strategy, then our result shows that level-k theory cannot be captured by the standard S5 system (or the partition model).

It is important to note that Theorem 4.3 is about the consistency of the theory, as formalized by $\left(\bigwedge_{\ell=0}^{k-1} \mathbf{B}_{e_\ell^i}(\mathbf{GR}_{i_\ell} \wedge g_{i_\ell})\right) \wedge \mathbf{B}_{e_k^i}(\mathbf{DF}_{i_k})$, not about the resulting solution. Indeed, if we apply the theorem to the undercutting game with R as the default choices,

[5]Note that it is not necessary for the Nash equilibrium to be unique. In fact, in this case, the default choices serve as a equilibrium selection device.

then level-k theory is consistent (under truth axiom) if and only if $k \leq 1$, even though the solution to level-3 and above is the Nash outcome.

These results give a precise sense that the level-k theory may be regarded as a model that deviates from full rationality—it violates logical consistency if the truth axiom is imposed for non-Nash behaviour. The source of this deviation comes from the finite bounds on interpersonal reasoning, k, but it is not clear whether one can escape from this inconsistency if the bound on interpersonal reasoning is removed, a topic we move to next.

5 Unbounded depths and Nash solution

The level-k theory assumes substantive rationality formulated by $\mathbf{GR}_i$, but it also postulates a cognitive bound k according to which at the deepest level, the kth layer of interpersonal thinking, the substantive rationality is replaced by the default actions. One advantage of the level-k theory is its finiteness, and the whole discourse is formulated in the finite system KD. However, this finiteness also necessitates the arbitrariness of the default choices at the deepest level.

Alternatively, here I consider the case where the players are not subject to a finite cognitive bound but can continue the interpersonal reasoning *ad infinitum*. Here I follow the approach taken by Hu and Kaneko (2014). In particular, I adopt the additional axioms there, which allow the players to obtain that *Nash equilibria* as the only viable options consistent with infinite regress of beliefs in substantive rationality. Moreover, when the game has a unique Nash equilibrium, the players can in fact derive the Nash equilibrium the *result* of inferences from interpersonal beliefs of the substantive rationality.

To pursue this approach, I introduce a new operator: for any pair of propositions, P_1 and P_2, and $i = 1, 2$, the operator $\mathbf{Ir}_i(P_1, P_2)$ is intended to express the infinite conjunction

$$\bigwedge_{\ell=0}^{\infty} \mathbf{B}_{e^i_\ell}(P_{i_\ell}). \tag{23}$$

Intuitively, the infinite conjunction $\mathbf{Ir}_i(P_1, P_2)$ captures the *infinite regress* of interpersonal beliefs regarding P_1 and P_2, with P_i occurs when the innermost belief operator is

$\mathbf{B}_i$, $i = 1, 2$. Moreover, the conjunction is from player i's perspective, as the outermost belief operator is always $\mathbf{B}_i$ in each of its elements. Hence, I shall call the proposition $\mathbf{Ir}_i(P_1, P_2)$ the infinite regress of P_1 and P_2 from i's perspective. The content of this infinite regress of beliefs, however, is given by $\mathbf{Ir}_i^o(P_1, P_2) = P_i \wedge \mathbf{Ir}_j(P_1, P_2)$. Indeed, from (23), $\mathbf{Ir}_i^o(P_1, P_2)$ corresponds to the following infinite conjunction:

$$P_i \wedge \left(\bigwedge_{\ell=0}^{\infty} \mathbf{B}_{e_\ell^j}(P_{j_\ell}) \right), \tag{24}$$

as each element in (23) is obtained by adding the belief operator $\mathbf{B}_i$ in front of each element in (24). Thus, intuitively, it should be the case that

$$\vdash \mathbf{Ir}_i(P_1, P_2) \equiv \mathbf{B}_i \left(\mathbf{Ir}_i^0(P_1, P_2) \right). \tag{25}$$

However, to handle proofs involving this new operator that contains infinity, the system KD needs to be extended.

Formally, the system KD is now extended in both its language and its logical axioms and inference rules, and the extended system is called IR. For the language, the set of propositions is now obtained by induction from (o)-(ii) in Section 2.1, plus the following step:

(iii) if P_1 and P_2 are propositions, so is $\mathbf{Ir}_i(P_1, P_2)$.

With the additional operator $\mathbf{Ir}_1$ and $\mathbf{Ir}_2$, the language in the system IR is still finite in the sense that all logical operations are applied to finite sets of propositions only, and the only "infinity" in the system is capture by the operators $\mathbf{Ir}_1$ and $\mathbf{Ir}_2$ directly without explicitly taking infinite conjunctions.[6] For logical axioms and inference rules, the system IR adds one more axiom and one more inference rule for the operators $\mathbf{Ir}_1$ and $\mathbf{Ir}_2$: for any propositions P_1 and P_2, and $i = 1, 2$, $j \neq i$,

Axiom $\mathbf{IRA}_i$: $\mathbf{Ir}_i(P_1, P_2) \Rightarrow (\mathbf{B}_i(P_i) \wedge \mathbf{B}_i(\mathbf{Ir}_j(P_1, P_2)))$.

The additional inference rule is given by: for any propositions D_1, D_2 and P_1, P_2,

Rule $\mathbf{IRI}_i$:

$$\frac{D_1 \Rightarrow (\mathbf{B}_1(P_1) \wedge \mathbf{B}_1(D_2)),\ D_2 \Rightarrow (\mathbf{B}_2(P_2) \wedge \mathbf{B}_2(D_1))}{D_i \Rightarrow \mathbf{Ir}_i(P_1, P_2)}.$$

[6] Alternatively, one could introduce an infinite language that allows for infinite conjunctions or other logical operations over infinite sets of propositions. The approach here is called the *fixed-point* approach and can be shown to be equivalent to the alternative approach with infinite languages. For a detailed discussion of the difference and comparisons, see Hu et al. (2019).

In terms of semantics, the definition of a Kripke frame remains the same as in Section 2.1, but a new evaluation rule for the operator $\mathbf{Ir}_i$ is needed:

- $\tau(\mathbf{Ir}_i(A_1, A_2), w) = \top$ iff $\tau(A_{i_k}, w_{k+1}) = \top$ for any alternating chain $\langle (w_0, i_0), ..., (w_k, i_k),\ w_{k+1} \rangle$ with $(w_0, i_0) = (w, i)$,

where $\langle (w_0, i_0), ..., (w_k, i_k), w_{k+1} \rangle \in (W \times \{1, 2\})^{k+1} \times W$ and $k \geq 0$ is an *alternating chain* if $i_{\ell-1} \neq i_{i_\ell}$ for $\ell = 1, ..., k$ and $w_{\ell-1} R_{i_{\ell-1}} w_\ell$ for $\ell = 1, ..., k+1$. The alternating structure corresponds to the set given by (23). The following theorem extends the completeness/soundness result to the system IR, and is adopted from Hu and Kaneko (2014).

Theorem 5.1 (Hu and Kaneko, 2014)**.** *For any formulae P in IR,*

$$\vdash P \text{ in IR if and only if } \models_M P \tag{26}$$

for all M in which W_1 and W_2 are serial.

Now I turn to the axioms of substantive rationality needed for the Nash solution. The first one is $\mathbf{GRN}_i$. I use this one instead of $\mathbf{GR}_i$ because the discourse requires the infinite regress all at once instead of layer-by-layer, as in the level-k theory. Indeed, in the level-k theory the interpersonal beliefs are anchored by the default choices at the deepest level, which pin down beliefs at shallower levels. In contrast, with infinite regress of beliefs, the beliefs are endogenously determined, and the infinite regress of $\mathbf{GRN}_i$ becomes itself the necessary condition. The sufficient condition for reaching a final decision, however, requires again to take the infinite regress all at once. Besides $\mathbf{GRN}_i$, I follow Hu and Kaneko (2014) and impose two additional axioms:

$$\mathbf{IA}_i: \quad \bigwedge_{a_i \in A_i} (\mathrm{D}_i(a_i) \Rightarrow \mathbf{B}_j \mathbf{B}_i(\mathrm{D}_i(a_i)))\,; \tag{27}$$

$$\mathbf{EP}_i: \quad \bigwedge_{a_i \in A_i} \left(\mathrm{D}_i(a_i) \Rightarrow \bigvee_{a_j \in A_j} \mathbf{B}_j(\mathrm{D}_j(a_j)) \right). \tag{28}$$

I call the axiom $\mathbf{IA}_i$ the *Interactive Axiom* for player i. The axiom captures the intuitive idea that player i assumes that his opponent has the symmetric logical abilities and shares the same substantive rationality. Hence, the axiom states that if

player i could reach the conclusion that a_i is indeed a possible final decision for himself, expressed by $\mathrm{D}_i(a_i)$, then player j can also reach this conclusion, expressed by $\mathbf{B}_j\mathbf{B}_i(\mathrm{D}_i(a_i))$. Thus, player i assumes that whatever he can infer as a possible final decision should not come as a surprise to his opponent. In this sense, this axiom is conceptually related to the truth axiom, which requires that the players' beliefs cannot be surprised by objective reality. But there is also a big difference: the interactive axiom does not assume anything about the objective reality but only about the opponent's ability to achieve the same conclusion given the beliefs.

Now let us turn to the derivation of the Nash derivation from the infinite regress of the substantive rationality modelled by the three axioms:

$$\mathbf{NN}_i \equiv \mathbf{GRN}_i \cap \mathbf{IA}_i \cap \mathbf{EP}_i, \tag{29}$$

where the first N stands for Nash and the second N for Necessity, as these all represent necessary conditions for a possible decision.

As mentioned, without imposing default choices, to determine possible final decisions it is necessary to have an infinite regress of interpersonal beliefs, that is, $\mathbf{Ir}_i(\mathbf{NN}_1, \mathbf{NN}_2)$. As in the level-$k$ theory, the goal here is to determine possible final decisions from this infinite regress. Different from the level-k theory, however, this cannot be done layer by layer, as in (21), but one has to take the infinite regress as a whole, since there is no anchor for the "deepest" layer of decisions.

To do so, I use a methodology similar to the characterization of the infinite regress operator $\mathbf{Ir}_i$ by the fixed-point argument. The idea is that to determine a possible final decision, we need to translate the requirements on $\mathrm{D}_i(a_i)$ in axioms $\mathbf{Ir}_i(\mathbf{NN}_1, \mathbf{NN}_2)$ into properties expressed by game-propositions, as I have characterized $\mathrm{D}_i(a_i)$ by $\widehat{D}^0(a_i)$, which is a game-proposition, in the level-k theory in Theorem 4.1. Now, the fixed-point approach then looks for candidate propositions indexed by actions, denoted by $E_i(a_i)$ and $E_j(a_j)$, that satisfies the requirements $\mathbf{Ir}_i(\mathbf{NN}_1, \mathbf{NN}_2)$.

Now, there will be two directions in the analysis. First I consider necessary conditions, that is, game-propositions that are implied by the axioms $\mathbf{Ir}_i(\mathbf{NN}_1, \mathbf{NN}_2)$. One such proposition is the following:

$$\hat{D}_i^*(a_i) \equiv \bigvee_{a_j \in A_j} \mathbf{Ir}_i^o(\mathrm{Bst}_i(a_i; a_j); \mathrm{Bst}_j(a_j; a_i)). \tag{30}$$

The following lemma shows that $\hat{D}_i^N(a_i)$ indeed follows from the axioms given by (29).

Lemma 5.1 (Hu and Kaneko, 2014)**.** *In IR, we have, for both $i = 1, 2$ and for each $a_i \in A_i$,*

$$\mathbf{Ir}_i(\mathbf{NN}_1, \mathbf{NN}_2) \vdash \mathbf{B}_i(D_i(a_i) \Rightarrow \hat{D}_i^*(a_i)). \tag{31}$$

Lemma 5.1 shows that $\hat{D}_i^*(a_i)$ is a necessary condition for a possible decision under infinite regress of substantive rationality given by **GRN**, **IA**, and **EP**. By the completeness and soundness theorem, Theorem 5.1, it also has implications for consistency, as the following theorem indicates.

Theorem 5.2. *Let G be a generic game satisfying (8) with formalized payoff (g_1, g_2).*

1. *The set $\{\mathbf{Ir}_i(\mathbf{NN}_i; \mathbf{NN}_j), \mathbf{Ir}_i(g_i; g_j)\}$ is consistent.*
2. *The set $\{\mathbf{Ir}_i(\mathbf{NN}_i; \mathbf{NN}_j), \mathbf{Ir}_i(g_i; g_j), D_i(a_i), \mathbf{B}_j(D_j(a_j))\}$ is consistent if and only if (a_i, a_j) is a Nash equilibrium in G.*

Theorem 5.2 then shows that the only actions that are consistent with the infinite regress of substantive rationality formulated here are those that constitute Nash equilibrium. Although similar to Theorem 4.3 in flavour, there are some key differences. First, there is no need of the truth axiom in Theorem 5.2. In fact, the result is not affected at all whether the truth axiom is imposed or not. Second, there is no need of the default actions to anchor the beliefs. Instead, the beliefs are endogenously determined when infinite regress is introduced.

Compared to Theorem 4.2, Theorem 5.2 falls short at giving concrete recommendation directly inferred from the infinite regress. It turns out that this is best one can do under the current formulation, and it captures the intuition that level-k theory can achieve a concrete recommendation because of the belief anchor rested on the default actions, while for the Nash equilibrium, multiplicity of equilibria prevents from a complete determination besides the consistency requirement for an equilibrium. Nash (1951) already anticipates this insight and proposes that Nash equilibrium can be *inferred* or solved, when the game is solvable. Below we show that, in that case, the Nash equilibrium can indeed be inferred in the same fashion as in Theorem 4.2.

Finally, I remark that when the game G has a unique Nash equilibrium, the axioms for substantive rationality thus formulated are sufficient to fully determine players'

predictions and actions. Note that the axioms, **GRN**, **IA**, and **EP**, are formulated in terms of necessary conditions, and hence they do not directly allow for the player to deem a possible decision as a recommendation. In Hu and Kaneko (2014) , they formulate parallel axioms for the other direction that essentially say that "if certain conditions are satisfied, then a_i is a possible final decision." With those axioms, they are able to show that

$$\mathbf{Ir}_i(\mathbf{NS}_i;\mathbf{NS}_j)\cup\{\mathbf{Ir}_i(\mathbf{NN}_i;\mathbf{NN}_j),\mathbf{Ir}_i(g_i;g_j)\}\vdash\mathbf{B}_i(\mathrm{D}_i(a_i)\equiv\hat{D}_i^*(a_i)), \tag{32}$$

where **NS** is the parallel axioms imposed. For self-containment I give these axioms in the Online Appendix. Thus, for games with unique Nash equilibrium, (32) then shows that player i can infer that the Nash strategy is a possible final decision from the assumptions. That is, from the substantive rationality formulated by (29), Nash equilibrium can be inferred from the infinite regress of rationality.

6 Concluding remarks

Conceptually, the truth axiom is closely related to rational expectation, which requires economic agents' subjective beliefs to be disciplined by the objective reality. The results reported here suggest that there is an intimate relationship between such assumption and the Nash equilibrium, at least under the assumption that economic agents only possess logically consistent beliefs. In stationary situations, default choices may be interpreted as repeatedly observed behaviour, and truth axiom and logical consistency seem sensible assumptions. These results then imply that it is reasonable to expect Nash equilibrium behaviour.

In contrast, in situations where agents are less familiar with, the truth axiom makes much less sense, and bounded interpersonal reasoning can be very subjective and the result above shows that level-k theory can be a coherent theory of behaviour when agents have bounded ability to conduct interpersonal inferences and can only reason up to a finite depth. However, if the agents can have unbounded interpersonal reasoning, logical consistency still requires equilibrium behaviour, but this may seem too demanding.

The results from the paper, however, also suggest that the most difficult case may arise in the intermediate case where agents are somehow familiar with the game situation so that the truth axiom is reasonable and default choices can be reasonably presumed, but it takes time for the agents to arrive at coherent strategies for play. The theory here predicts a tension between non-equilibrium behaviour and logical inconsistency that may lead to further adjustments in agents' subjective beliefs and behaviour. However, further research is needed to understand how such adjustment would occur, both theoretically and empirically.

Appendix A: Proofs of lemmas and theorems

Before the proof of Theorem 4.1, I first list a few useful lemmas. Let D be a given formula. Let D be a formula and A a subformula of D, and let C be another formula. We use $D\langle A/C\rangle$ to denote the resulting formula by replacing all occurrences of A by C.

Lemma 6.1 (Substitutability lemma)**.** *Let A, C, D be formulas. Suppose that A does not occur within the scope of any occurrence of $\mathbf{B}_i(\cdot)$ for either $i = 1, 2$. Then, for any formula, E,*

$$\text{if } E \vdash A \equiv C \text{ and } E \vdash D, \text{ then } E \vdash D\langle A/C\rangle. \tag{33}$$

Proof. By the Completeness Theorem, I prove the following statement:

$$(M, w) \vDash E \text{ implies that } (M, w) \vDash D'\langle A/C\rangle \text{ if and only if } (M, w) \vDash D'\langle A/C\rangle. \tag{34}$$

for all models M and any subformula D' containing A. I prove this by induction on D' as a subformula of D w.r.t. the connectives, beginning with the induction base where $D' = A$. The induction base follows immediately. By induction, it is sufficient to show that, assuming that (34) holds for D' and D'', it also holds for $F = \neg D'$, $F = D' \wedge D''$, and $F = D' \Rightarrow D''$. These then follow immediately from the evaluation rules for τ. Crucially, note that there is no need to consider the case where $F = \mathbf{B}_i(D')$ as A does not occur within the scope of any occurrence of $\mathbf{B}_i(\cdot)$ for either $i = 1, 2$. □

Lemma 6.2. *Let Φ be a finite set of formulas. Then,*

$$\vdash \mathbf{B}_i(\wedge\Phi) \equiv \left(\bigwedge_{A \in \Phi} \mathbf{B}_i(A) \right). \tag{35}$$

Proof of Theorem 4.1

I prove (21) by induction on $m = k, k-1, ..., 1$. The induction base is (21) for $m = k$, where the statement becomes

$$\mathbf{B}_{i_k}(\mathbf{DF}_{i_k}) \vdash \mathbf{B}_{i_k}\left(\bigwedge_{a_{i_k} \in A_{i_k}} (\mathrm{D}_{i_k}(a_{i_k}) \equiv \hat{D}^k_{i_k}(a_{i_k}))\right).$$

This then follows directly from (15), Nec, and K.

Now, suppose, by induction hypothesis, that (34) holds for $m \leq k$, that is,

$$\left(\bigwedge_{\ell=m}^{k-1} \mathbf{B}_{e^{i_m}_{\ell-m}}(\mathbf{GR}_{i_\ell})\right) \wedge \mathbf{B}_{e^{i_m}_{k-m}}(\mathbf{DF}_{i_k}) \vdash \mathbf{B}_{i_m}\left(\bigwedge_{a_{i_m} \in A_{i_m}} (\mathrm{D}_{i_m}(a_{i_m}) \equiv \widehat{D}^m_{i_m}(a_{i_m}))\right),$$

which implies that, for all $a_{i_m} \in A_{i_m}$,

$$\left(\bigwedge_{\ell=m}^{k-1} \mathbf{B}_{e^{i_m}_{\ell-m}}(\mathbf{GR}_{i_\ell})\right) \wedge \mathbf{B}_{e^{i_m}_{k-m}}(\mathbf{DF}_{i_k}) \vdash \mathbf{B}_{i_m}(\mathrm{D}_{i_m}(a_{i_m})) \equiv \mathbf{B}_{i_m}(\widehat{D}^m_{i_m}(a_{i_m})). \tag{36}$$

Now, by (11), we have, for all $a_{i_{m-1}} \in A_{i_{m-1}}$,

$$\mathbf{GR}_{i_{m-1}} \vdash \mathrm{D}_{i_{m-1}}(a_{i_{m-1}}) \equiv \bigwedge_{a_{i_m} \in A_{i_m}} \left(\mathbf{B}_{i_m}(\mathrm{D}_{i_m}(a_{i_m})) \Rightarrow \mathrm{Bst}_{i_{m-1}}(a_{i_{m-1}}; a_{i_m})\right). \tag{37}$$

Now, take

$$E = \mathbf{GR}_{i_{m-1}} \wedge \left(\bigwedge_{\ell=m}^{k-1} \mathbf{B}_{e^{i_m}_{\ell-m}}(\mathbf{GR}_{i_\ell})\right) \wedge \mathbf{B}_{e^{i_m}_{k-m}}(\mathbf{DF}_{i_k}),$$

and

$$A = \mathbf{B}_{i_m}(\mathrm{D}_{i_m}(a_{i_m})) \text{ and } C = \mathbf{B}_{i_m}(\widehat{D}^m_{i_m}(a_{i_m})),$$

and D to be the right-side of (37), by Lemma 6.1 (note that we need to use it repeatedly, one a_{i_m} at a time), we have

$$E \vdash \mathrm{D}_{i_{m-1}}(a_{i_{m-1}}) \equiv \bigwedge_{a_{i_m} \in A_{i_m}} \left(\mathbf{B}_{i_m}(\widehat{D}^m_{i_m}(a_{i_m})) \Rightarrow \mathrm{Bst}_{i_{m-1}}(a_{i_{m-1}}; a_{i_m})\right).$$

Since $\widehat{D}^{m-1}_{i_{m-1}}(a_{i_{m-1}}) = \bigwedge_{a_{i_m} \in A_{i_m}} \left(\mathbf{B}_{i_m}(\widehat{D}_{i_m}(a_{i_m})) \Rightarrow \mathrm{Bst}_{i_{m-1}}(a_{i_{m-1}}; a_{i_m})\right)$ by (19), it follows that, for all $a_{i_{m-1}} \in A_{i_{m-1}}$,

$$\mathbf{GR}_{i_{m-1}} \wedge \left(\bigwedge_{\ell=m}^{k-1} \mathbf{B}_{e^{i_m}_{\ell-m}}(\mathbf{GR}_{i_\ell})\right) \wedge \mathbf{B}_{e^{i_m}_{k-m}}(\mathbf{DF}_{i_k}) \vdash \mathrm{D}_{i_{m-1}}(a_{i_{m-1}}) \equiv \widehat{D}^{m-1}_{i_{m-1}}(a_{i_{m-1}}).$$

Then, applying Nec, $\wedge$-rule, and K, we obtain

$$\mathbf{B}_{i_{m-1}}(E) \vdash \mathbf{B}_{i_{m-1}}\left(\bigwedge_{a_{i_{m-1}} \in A_{i_{m-1}}} (\mathrm{D}_{i_{m-1}}(a_{i_{m-1}}) \equiv \widehat{D}^{m-1}_{i_{m-1}}(a_{i_{m-1}})\right).$$

Finally, note that

$$\vdash \mathbf{B}_{i_{m-1}}(E) \equiv \left(\bigwedge_{\ell=m-1}^{k-1} \mathbf{B}_{e^{i_{m-1}}_{\ell-m+1}}(\mathbf{GR}_{i_\ell})\right) \wedge \mathbf{B}_{e^{i_{m-1}}_{k-m+1}}(\mathbf{DF}_{i_k}),$$

Note: need a lemma for this. and hence we have proved the assertion for the induction step.■

Proof of Theorem 4.2

Characterization. First I show by induction on m, from $m = k-1$ to $m = 0$, that

$$\bigwedge_{\ell=m}^{k-1} \mathbf{B}_{e^{i_m}_{\ell-m}}(g_{i_\ell}) \vdash \mathbf{B}_{i_m}(\widehat{D}^m_{i_m}(a^{k-m}_{i_m})) \wedge \left(\bigwedge_{a_{i_m} \neq a^{k-m}_{i_m}} \mathbf{B}_{i_m}(\neg \widehat{D}^m_{i_m}(a_{i_m}))\right), \tag{38}$$

where $a^{k-m}_{i_m}$ for $m = 1, .., k$ is given by (12)-(14). Note that the desired result follows then immediately from (20) and (38) with $m = 0$.

Now consider (38) for $m = k-1$, which reads

$$\mathbf{B}_{i_{k-1}}(g_{i_{k-1}}) \vdash \mathbf{B}_{i_{k-1}}(\widehat{D}^{k-1}_{i_{k-1}}(a^1_{i_{k-1}})) \wedge \left(\bigwedge_{a_{i_{k-1}} \neq a^1_{i_{k-1}}} \mathbf{B}_{i_{k-1}}(\neg \widehat{D}^{k-1}_{i_{k-1}}(a_{i_{k-1}}))\right),$$

which, by (35), is in turn equivalent to

$$\mathbf{B}_{i_{k-1}}(g_{i_{k-1}}) \vdash \mathbf{B}_{i_{k-1}}\left(\widehat{D}^{k-1}_{i_{k-1}}(a^1_{i_{k-1}}) \wedge \left(\bigwedge_{a_{i_{k-1}} \neq a^1_{i_{k-1}}} \neg \widehat{D}^{k-1}_{i_{k-1}}(a_{i_{k-1}})\right)\right),$$

but by the Scope Lemma (Theorem 2.2), we only need to show

$$g_{i_{k-1}} \vdash \widehat{D}^{k-1}_{i_{k-1}}(a^1_{i_{k-1}}) \wedge \left(\bigwedge_{a_{i_{k-1}} \neq a^1_{i_{k-1}}} \neg \widehat{D}^{k-1}_{i_{k-1}}(a_{i_{k-1}})\right). \tag{39}$$

Now, by Nec and the definition of $\mathrm{L}^0_{i_k}(a_{i_k})$, we have

$$\vdash \mathbf{B}_{i_k}(\mathrm{L}^0_{i_k}(\bar{a}_{i_k})),\ \vdash \neg\mathbf{B}_{i_k}(\mathrm{L}^0_{i_k}(a_{i_k})) \text{ for all } a_{i_k} \neq \bar{a}_{i_k}.$$

Since $g_{i_{k-1}} \vdash \mathrm{Bst}_{i_{k-1}}(a^1_{i_{k-1}}; \bar{a}_{i_k})$ and $g_{i_{k-1}} \vdash \neg\mathrm{Bst}_{i_{k-1}}(a_{i_{k-1}}; \bar{a}_{i_k})$ for all $a_{i_{k-1}} \neq a^1_{i_{k-1}}$, we have

$$g_{i_{k-1}} \vdash \mathbf{B}_{i_k}(\mathrm{L}^0_{i_k}(\bar{a}_{i_k})) \Rightarrow \mathrm{Bst}_{i_{k-1}}(a^1_{i_{k-1}}; \bar{a}_{i_k}),$$

and for any $a_{i_k} \neq \bar{a}_{i_k}$,

$$g_{i_{k-1}} \vdash \mathbf{B}_{i_k}(\mathrm{L}^0_{i_k}(a_{i_k})) \Rightarrow \mathrm{Bst}_{i_{k-1}}(a^1_{i_{k-1}}; a_{i_k}),$$

where the first inference holds because $g_{i_{k-1}} \vdash \mathrm{Bst}_{i_{k-1}}(a^1_{i_{k-1}}; \bar{a}_{i_k})$ and the second holds because $\vdash \neg\mathbf{B}_{i_k}(\mathrm{L}^0_{i_k}(a_{i_k}))$. This implies that $g_{i_{k-1}} \vdash \widehat{D}^{k-1}_{i_{k-1}}(a^1_{i_{k-1}})$. Similarly, for any $a_{i_{k-1}} \neq a^1_{i_{k-1}}$, by (8),

$$g_{i_{k-1}} \vdash \mathbf{B}_{i_k}(\mathrm{L}^0_{i_k}(\bar{a}_{i_k})) \wedge \neg\mathrm{Bst}_{i_{k-1}}(a_{i_{k-1}}; \bar{a}_{i_k}),$$

which implies that $g_{i_{k-1}} \vdash \neg\widehat{D}^{k-1}_{i_{k-1}}(a_{i_{k-1}})$. This proves (39).

Now, by the induction hypothesis, suppose that (38) holds for some $m \leq k-1$ and now consider the case with $m-1$. By (35) and the Scope Lemma, we only need to show

$$g_{i_{m-1}} \wedge \left(\bigwedge_{\ell=m}^{k-1} \mathbf{B}_{e^{i_m}_{\ell-m}}(g_{i_\ell})\right) \vdash \widehat{D}^{m-1}_{i_{m-1}}(a^{k-m+1}_{i_{m-1}}) \wedge \left(\bigwedge_{a_{i_{m-1}} \neq a^{k-m+1}_{i_{m-1}}} \neg\widehat{D}^{m-1}_{i_{m-1}}(a_{i_{m-1}})\right). \quad (40)$$

Now, by (8), we have

$$g_{i_{m-1}} \vdash \mathrm{Bst}_{i_{m-1}}(a^{k-m+1}_{i_{m-1}}; a^{k-m}_{i_m}) \text{ and } g_{i_{m-1}} \vdash \neg\mathrm{Bst}_{i_{m-1}}(a_{i_{m-1}}; a^{k-m}_{i_m}) \text{ for any } a_{i_{m-1}} \neq a^{k-m+1}_{i_{m-1}}.$$

Thus, we have

$$g_{i_{m-1}} \wedge \left(\bigwedge_{\ell=m}^{k-1} \mathbf{B}_{e^{i_m}_{\ell-m}}(g_{i_\ell})\right) \vdash \mathbf{B}_{i_m}(\widehat{D}^m_{i_m}(a^{k-m}_{i_m})) \Rightarrow \mathrm{Bst}_{i_{m-1}}(a^{k-m+1}_{i_{m-1}}; a^{k-m}_{i_m})$$

and that for any $a_{i_m} \neq a^{k-m}_{i_m}$,

$$g_{i_{m-1}} \wedge \left(\bigwedge_{\ell=m}^{k-1} \mathbf{B}_{e^{i_m}_{\ell-m}}(g_{i_\ell})\right) \vdash \mathbf{B}_{i_m}(\widehat{D}^m_{i_m}(a_{i_m})) \Rightarrow \mathrm{Bst}_{i_{m-1}}(a^{k-m+1}_{i_{m-1}}; a_{i_m}),$$

where the second inference follows from the fact that by the induction hypothesis, (38),

$$g_{i_{m-1}} \wedge \left(\bigwedge_{\ell=m}^{k-1} \mathbf{B}_{e^{i_m}_{\ell-m}}(g_{i_\ell}) \right) \vdash \neg \mathbf{B}_{i_m} \widehat{D}^m_{i_m}(a_{i_m}).$$

Similarly, for any $a_{i_{m-1}} \neq a^{k-m+1}_{i_{m-1}}$,

$$g_{i_{m-1}} \wedge \left(\bigwedge_{\ell=m}^{k-1} \mathbf{B}_{e^{i_m}_{\ell-m}}(g_{i_\ell}) \right) \vdash \mathbf{B}_{i_m}(\widehat{D}^m_{i_m}(a^{k-m}_{i_m})) \wedge \neg \mathrm{Bst}_{i_{m-1}}(a^{k-m+1}_{i_{m-1}}; a^{k-m}_{i_m}).$$

This proves (40).

Consistency of the assumptions. By Theorem 2.1, it is sufficient to construct a model M for $\left(\bigwedge_{\ell=0}^{k-1} \mathbf{B}_{e^i_\ell}(\mathbf{GR}_{i_\ell} \wedge g_{i_\ell})\right) \wedge \mathbf{B}_{e^i_k}(\mathbf{DF}_{i_k})$. Let $\mathbf{a}_i = (a^1_i, a^2_j, ..., a^k_i)$ be the sequence constructed in (12)-(14). The construction is as follows:

$$\begin{aligned}
W &= \{w_1, ..., w_{k+1}\};\\
R_i &= \{(w_{\nu+1}, w_\nu) : \nu = 1, ..., k\} \cup \{(w_1, w_1)\};\\
R_j &= \{(w_{\nu+1}, w_\nu) : \nu = 1, ..., k\} \cup \{(w_1, w_1)\};\\
\tau(w_\nu, \mathrm{Pr}_i(a; a')) &= \top \text{ iff } a \succeq a' \text{ for all } \nu = 1, ..., k+1;\\
\tau(w_1, \mathrm{D}_{i_k}(a_{i_k})) &= \top \text{ iff } a_{i_k} = \bar{a}_{i_k},\ \tau(w_1, \mathrm{D}_{i_{k-1}}(a_{i_{k-1}})) = \top \text{ iff } a_{i_{k-1}} = a^1_{i_{k-1}};\\
&\text{for all } \nu = 2, ..., k,\\
\tau(w_\nu, \mathrm{D}_{i_{k-\nu}}(a_{i_{k-\nu}})) &= \top \text{ iff } a_{i_{k-\nu}} = a^\nu_{i_{k-\nu}},\ \tau(w_\nu, \mathrm{D}_{i_{k-\nu+1}}(a_{i_{k-\nu+1}})) = \bot \text{ for all } a_{i_{k-\nu-1}};\\
\tau(w_{k+1}, \mathrm{D}_i(a_i)) &= \bot = \tau(w_{k+1}, \mathrm{D}_j(a_j)) \text{ for all } (a_i, a_j) \in A_i \times A_j.
\end{aligned} \tag{41}$$

We claim that, for $\nu = 1, ..., k$,

$$(M, w_\nu) \models (\mathbf{GR}_{i_{k-\nu}} \wedge g_{i_{k-\nu}}) \wedge \left(\bigwedge_{\ell=k-(\nu-1)}^{k-1} \mathbf{B}_{e^{i_{k-(\nu-1)}}_{\ell-k+(\nu-1)}}(\mathbf{GR}_{i_\ell} \wedge g_{i_\ell}) \right) \wedge \mathbf{B}_{e^{i_{k-(\nu-1)}}_{\nu-1}}(\mathbf{DF}_{i_k}). \tag{42}$$

Since $(w_{k+1}, u) \in R_i$ holds only for $u = w_k$, it follows from (42) with $\nu = k$ that

$$(M, w_{k+1}) \models \mathbf{B}_{i_0} \left(\mathbf{GR}_{i_0} \wedge g_{i_0} \wedge \left(\bigwedge_{\ell=1}^{k-1} \mathbf{B}_{e^{i_1}_{\ell-1}}(\mathbf{GR}_{i_\ell} \wedge g_{i_\ell}) \right) \wedge \mathbf{B}_{e^{i_1}_{k-1}}(\mathbf{DF}_{i_k}) \right),$$

which, by Lemma 6.2 and the Completeness Theorem, implies that

$$(M, w_{k+1}) \models \left(\bigwedge_{\ell=0}^{k-1} \mathbf{B}_{e^i_\ell}(\mathbf{GR}_{i_\ell} \wedge g_{i_\ell}) \right) \wedge \mathbf{B}_{e^i_k}(\mathbf{DF}_{i_k}).$$

Now we show (42) by induction on ν, beginning with $\nu = 1$. By construction, we have

$$(M, w_1) \models \mathbf{B}_{i_k}(\mathrm{D}_{i_k}(\bar{a}_{i_k})) \wedge \mathrm{D}_{i_{k-1}}(a^1_{i_{k-1}}) \wedge \mathrm{Bst}_{i_{k-1}}(a^1_{i_{k-1}}; \bar{a}_{i_k}),$$

and that, for any $a_{i_k} \neq \bar{a}_{i_k}$ and any $a_{i_{k-1}} \neq a^1_{i_{k-1}}$,

$$(M, w_1) \models \neg\mathbf{B}_{i_k}(\mathrm{D}_{i_k}(a_{i_k})) \wedge \neg\mathrm{D}_{i_{k-1}}(a_{i_{k-1}}).$$

It follows that

$$(M, w_1) \models (\mathbf{GR}_{i_{k-1}} \wedge g_{i_{k-1}}) \wedge \mathbf{B}_{e^{i_k}_0}(\mathbf{DF}_{i_k}),$$

that is, (42) for $\nu = 1$.

Now, by induction hypothesis, suppose that (42) holds for ν and consider the case for $\nu + 1$. Since $(w_{\nu+1}, u) \in R_{i_{k-\nu}}$ holds only for $u = w_\nu$, it follows from (42) with ν that

$$(M, w_{\nu+1}) \models \mathbf{B}_{i_{k-\nu}} \left(\mathbf{GR}_{i_{k-\nu}} \wedge g_{i_{k-\nu}} \wedge \left(\bigwedge_{\ell=k-(\nu-1)}^{k-1} \mathbf{B}_{e^{i_{k-(\nu-1)}}_{\ell-k+(\nu-1)}}(\mathbf{GR}_{i_\ell} \wedge g_{i_\ell}) \right) \wedge \mathbf{B}_{e^{i_{k-(\nu-1)}}_{\nu-1}}(\mathbf{DF}_{i_k}) \right),$$

which, by Lemma 6.2 and the Completeness Theorem, implies that

$$(M, w_{\nu+1}) \models \left(\bigwedge_{\ell=k-\nu}^{k-1} \mathbf{B}_{e^{i_{k-\nu}}_{\ell-k+\nu}}(\mathbf{GR}_{i_\ell} \wedge g_{i_\ell}) \right) \wedge \mathbf{B}_{e^{i_{k-\nu}}_{\nu}}(\mathbf{DF}_{i_k}).$$

Thus, we only need to show that $(M, w_{\nu+1}) \models \mathbf{GR}_{i_{k-(\nu+1)}}$. By construction, we have

$$(M, w_{\nu+1}) \models \mathbf{B}_{i_{k-\nu}}(\mathrm{D}_{i_{k-\nu}}(a^\nu_{i_{k-\nu}})) \wedge \mathrm{D}_{i_{k-(\nu+1)}}(a^{\nu+1}_{i_{k-(\nu+1)}}) \wedge \mathrm{Bst}_{i_{k-(\nu+1)}}(a^{\nu+1}_{i_{k-(\nu+1)}}; a^\nu_{i_{k-\nu}}),$$

and that, for any $a_{i_{k-\nu}} \neq a^\nu_{i_{k-\nu}}$ and any $a_{i_{k-(\nu+1)}} \neq a^{\nu+1}_{i_{k-(\nu+1)}}$,

$$(M, w_1) \models \neg\mathbf{B}_{i_{k-\nu}}(\mathrm{D}_{i_{k-\nu}}(a_{i_{k-\nu}})) \wedge \neg\mathrm{D}_{i_{k-(\nu+1)}}(a_{i_{k-(\nu+1)}}).$$

This implies that $(M, w_{\nu+1}) \models \mathbf{GR}_{i_{k-(\nu+1)}}$ and completes the proof.

Proof of Theorem 4.3

Consistency when (a^k_i, a^{k-1}_j) is a Nash equilibrium.

In this case, we build a model M in which $W = \{w\}$ and $R_i = R_j = \{(w, w)\}$. Since $(a^*_i, a^*_j) = (a^k_i, a^{k-1}_j)$ is a Nash equilibrium, it follows from (8) that $a^{k-2t}_{i_{2t}}$ is constant

and $a^{k-2t-1}_{i_{2t+1}}$ is constant for $t = 1, 2, ...\lfloor k/2 \rfloor$. Moreover, $\bar{a}_{i_k} = a^k_i$ for odd and $\bar{a}_{i_k} = a^{k-1}_j$ for k even. The truth evaluations of the atomic formulas are given by

$$\begin{aligned} \tau(w, \mathrm{Pr}_i(a; a')) &= \top \text{ iff } a \succeq a'; \\ \tau(w, \mathrm{D}_i(a_i)) &= \top \text{ iff } a_i = a^*_i, \\ \tau(w, \mathrm{D}_j(a_j)) &= \top \text{ iff } a_j = a^*_j; \end{aligned} \tag{43}$$

Then, it follows that

$$(M, w) \models \mathbf{DF}_{i_k} \wedge \mathbf{GR}_i \wedge \mathbf{GR}_j \wedge \mathbf{g}.$$

Consistency for $k = 2$

Inconsistency of Level-k theory

Since adding Axiom T still preserves all the theorems, (21) still holds under Axiom T. Assume that $k \geq 2$ is even; the other case is similar. Since (a^k_i, a^{k-1}_j) is not a Nash equilibrium, it follows from (8) that $\bar{a}_i \neq a^2_i$.

Now, by Axiom T,

$$\left(\bigwedge_{\ell=0}^{k-1} \mathbf{B}_{e^i_\ell}(\mathbf{GR}_{i_\ell} \wedge g_{i_\ell}) \right) \wedge \mathbf{B}_{e^i_k}(\mathbf{DF}_{i_k}) \vdash \mathbf{B}_i\mathbf{B}_j(\mathbf{GR}_j) \wedge \mathbf{B}_i(\mathbf{GR}_i) \wedge \mathbf{B}_i\mathbf{B}_j\mathbf{B}_i(\mathbf{DF}_i),$$

and

$$\left(\bigwedge_{\ell=0}^{k-1} \mathbf{B}_{e^i_\ell}(\mathbf{GR}_{i_\ell} \wedge g_{i_\ell}) \right) \wedge \mathbf{B}_{e^i_k}(\mathbf{DF}_{i_k}) \vdash \mathbf{DF}_i.$$

Then, by (21),

$$\mathbf{B}_i\mathbf{B}_j(\mathbf{GR}_j) \wedge \mathbf{B}_i(\mathbf{GR}_i) \wedge \mathbf{B}_i\mathbf{B}_j\mathbf{B}_i(\mathbf{DF}_i) \vdash \mathbf{B}_i \left(\bigwedge_{a_i \in A_i} (\mathrm{D}_i(a_i) \equiv \widehat{D}^{k-2}_i(a_i)) \right).$$

and

$$\mathbf{DF}_i \vdash \mathrm{D}_i(\bar{a}_i) \wedge \neg D_i(a^2_i).$$

Now, by the proof of Theorem 4.2,

$$\mathbf{B}_i(g_i) \wedge \mathbf{B}_i\mathbf{B}_j(g_j) \vdash \mathbf{B}_i(\widehat{D}^{k-2}_i(a^2_i) \wedge \neg \widehat{D}^{k-2}_i(\bar{a}_i)).$$

Finally, using Axiom T, we obtain

$$\left(\bigwedge_{\ell=0}^{k-1} \mathbf{B}_{e^i_\ell}(\mathbf{GR}_{i_\ell} \wedge g_{i_\ell}) \right) \wedge \mathbf{B}_{e^i_k}(\mathbf{DF}_{i_k}) \vdash \neg D_i(a^2_i) \wedge D_i(a^2_i),$$

which shows that the assumptions are inconsistent.

Proof of Theorem 5.1

Lemma 6.3. ***(1)****:* $\vdash \mathbf{Ir}_i(P_1, P_2) \equiv \mathbf{B}_i\,(\mathbf{Ir}_i^o(P_1, P_2))$*;*

(2)*:* *if* $\vdash P_k$ *for* $k = 1, 2,$ *then* $\vdash \mathbf{Ir}_i(P_1, P_2)$*;*

(3)*:* $\vdash \mathbf{Ir}_i(P_1, P_2) \Rightarrow \mathbf{Ir}_i\,(\mathbf{Ir}_1^o(P_1, P_2), \mathbf{Ir}_2^o(P_1, P_2))\,;$

(4)*:* $\vdash \mathbf{Ir}_i(P_1 \Rightarrow Q_1, P_2 \Rightarrow Q_2) \wedge \mathbf{Ir}_i(P_1, P_2) \Rightarrow \mathbf{Ir}_i(Q_1, Q_2),$ *equivalently,*
$\vdash \mathbf{Ir}_i(P_1 \Rightarrow Q_1, P_2 \Rightarrow Q_2) \Rightarrow (\mathbf{Ir}_i(P_1, P_2) \Rightarrow \mathbf{Ir}_i(Q_1, Q_2))$*;*

(5)*:* $\vdash \mathbf{Ir}_i(P_1, P_2) \wedge \mathbf{Ir}_i[Q_1, Q_2] \equiv \mathbf{Ir}_i(P_1 \wedge Q_1, P_2 \wedge Q_2).$