

Understanding the Brandenburger-Keisler Belief Paradox

Eric Pacuit

Institute of Logic, Language and Information

University of Amsterdam

`epacuit@staff.science.uva.nl`

`staff.science.uva.nl/~epacuit`

March 15, 2006

Abstract

Adam Brandenburger and H. Jerome Keisler have recently discovered a two person Russell-style paradox. They show that the following configurations of beliefs is impossible: *Ann believes that Bob assumes that Ann believes that Bob's assumption is wrong*. In [7] a modal logic interpretation of this paradox is proposed. The idea is to introduce two modal operators intended to represent the agents' *beliefs* and *assumptions*. The goal of this paper is to take this analysis further and study this paradox from the point of view of a modal logician. In particular, we show that the paradox can be seen as a theorem of an appropriate hybrid logic. Furthermore, we propose a sound and complete axiomatization of a modal logic with belief and assumption modal operators, a question left open in [7].

1 Introduction

In their textbook, Osborne and Rubinstein describe game theory as “a bag of analytical tools designed to help us understand the phenomena that we observe when decision-makers interact” ([15] page 1). They go on to say that one of the basic assumptions of game theory is that when agents make decisions, they take into account “their knowledge or expectations of *other* decision-makers' behavior (they *reason strategically*).” In other words, when agents involved in a multi-agent interactive situation are making decisions about what action to perform

next, that decision is influenced by what actions they *expect* the other agents will perform. This assumption leads very naturally to questions about what agents believe about the other agents' beliefs.

This observation has prompted a number of game theorists to propose that the basic models of game theory (extensive games forms and normal game forms) be extended to include a representation of the agents' beliefs (see [1, 17, 5, 6, 13] for a discussion of the relevant literature). Essentially the idea is that when describing a strategic interactive situation part of that description should include the agents' beliefs about the relevant ground (non-epistemic) facts, beliefs about the other agents' beliefs about these ground facts, beliefs about the other agents' beliefs about the other agents' beliefs about these ground facts, and so on. Early on in 1967, John Harsanyi [12] developed an elegant formal model which can be used to represent the epistemic state of the agent in a game theoretic situation¹. The idea is that each agent could be any one of a number of different **types**, where a type is intended to represent an *infinite hierarchy of beliefs*, i.e., the agent's first-order beliefs about the strategies of the other agents, second-order beliefs about the other agents' first-order beliefs, third-order beliefs about the other agents' second-order beliefs, and so on. Thus the problem of adding beliefs to the basic models of game theory reduces to finding an appropriate collection of possible types for each agent.

Adam Brandenburger and H. Jerome Keisler have recently discovered a Russell-style paradox lurking in the background of the above discussion. In [7], they show that the following configurations of beliefs is impossible: *Ann believes that Bob assumes² that Ann believes that Bob's assumption is wrong*. This suggests that it may not always be possible to find a type space to represent certain configurations of beliefs.

In [7] a modal logic interpretation of the paradox is proposed. The idea is to introduce two modal operators intended to represent the agents' *beliefs* and *assumptions*. The goal of this paper is to take this analysis further and study this paradox from the point of view of a modal logician. In particular, we show that the paradox can be seen as a theorem of an appropriate hybrid logic³. Furthermore, we propose a sound and complete axiomatization of a modal logic with belief and assumption modal operators, a question left open in [7].

¹Harsanyi's original motivation was to study games of *incomplete* information, i.e., games in which the agents are uncertain about the structure of the game.

²An assumption is a belief that implies all other beliefs. It is shown in [7] that it is crucial the statement be about "one particular belief of Bob and all of Ann's beliefs".

³Hybrid logic is a modal logic with distinguished propositional variables called *nominals* that are used to name each world in a Kripke structure. See [4] for more information.

2 The Paradox

In [7], Brandenburger and Keisler introduce the following two person Russell-style paradox. The statement of the paradox involves two concepts: beliefs and assumptions. An assumption is assumed to be a strongest belief. We will say more about the interpretation of an assumption below. Suppose there are two players, Ann and Bob, and consider the following description of beliefs.

Ann believes that Bob assumes that Ann believes that Bob's assumption is wrong.

A paradox arises when one asks the question

Does Ann believe that Bob's assumption is wrong?

Suppose that answer to the above question is 'yes'. Then according to Ann, Bob's assumption is wrong. But, according to Ann, Bob's assumption is *Ann believes that Bob's assumption is wrong*. However, since the answer to the above question is 'yes', Ann believes that this assumption is *correct*. So Ann does not believe that Bob's assumption is wrong. Therefore, the answer to the above question must be 'no'. Thus, it is not the case that Ann believes that Bob's assumption is wrong. Hence Ann believes Bob's assumption is correct. That is, it is correct that Ann believes that Bob's assumption is wrong. So, the answer must have been yes. This is a contradiction.

Just as Russell's paradox suggests that not every collection can constitute a set, the Brandenburger-Keisler paradox suggests that not every description of beliefs can be "represented". This can be made precise by introducing a **belief model** intended to represent each agent's beliefs about the other agent's beliefs. Formally, a belief model is a two-sorted structure intended to represent the beliefs each agent has about the other agent's beliefs. Each sort (one for each agent) is intended to represent a possible epistemic state of an agent. Let x, x', x_1, x_2, \dots denote "Ann states" and y, y', y_1, y_2, \dots denote "Bob states". Let W^a and W^b denote the set of Ann and Bob states respectively. For simplicity, we assume these sets are disjoint. Thus the collection of states is the set $W = W^a \cup W^b$. The second component of a belief model is a pair of relations (one for each agent), denoted P^a and P^b . The intended interpretation of xP^ay , where $x \in W^a$ and $y \in W^b$, is that in state x , Ann considers y possible. Similarly for Bob. Since the object of Ann's beliefs is a statement about Bob's beliefs, a **language for Ann**, denoted \mathcal{L}^a , is any collection of subsets of W^b . Similarly for Bob. A **language** is the (disjoint) union of \mathcal{L}^a and \mathcal{L}^b . For example, the **power set language** is

the language $\mathcal{L} = 2^{W^b} \cup 2^{W^a}$. Given a proposition $Y \in \mathcal{L}^a$ (i.e., a collection of Bob states), Ann is said to **believe** Y at state x provided the set of states that Ann considers possible at x is a subset of Y . An **assumption** is defined to be a strongest belief. That is, given a set $Y \in \mathcal{L}^b$, Ann is said to **assume** Y if Y equals the set of states that Ann considers possible. Similarly for Bob.

We can now be more explicit about what it means to say that not every configurations of beliefs can be represented. Say that a language \mathcal{L} is **complete** for a belief model if every statement in a player's language which is possible (i.e., true for some states) can be *assumed* by the player. It is not too difficult to construct an argument (using Cantor's Theorem) that no model is complete for its powerset language. The main result of [7] is that the same is true for the first-order language. Let the **first-order language** of a belief model be the collection of first-order definable subsets of $W = W^a \cup W^b$.

Theorem 1 (Brandenburger and Keisler [7]) *No belief model is complete for its first-order language.*

The proof proceeds by formalizing the above paradox. The next section examines this proof in detail.

3 Using Modal Logic

In [7], a modal logic version of Theorem 1 is proposed. The idea is to think of the interactive belief models defined in the previous section as Kripke structures. The approach we take in this section is to use *neighborhood models*. Neighborhood models are a generalization of the standard Kripke, or relational, semantics for modal logic invented by Scott and Montague (independently in [16] and [14]). See [8] for a discussion of the basic results concerning neighborhood models and the logics that correspond to them. A neighborhood model consists of a nonempty set of states, a valuation function and a neighborhood function that maps states to sets of sets of states, i.e., if W is the set of states then a neighborhood function is a map from W to 2^{2^W} . Given a state $w \in W$, a modal formula $\Box\phi$ is said to be true at state w provided the truth set of ϕ (the set of all states satisfying ϕ) is an element of the neighborhood of w . Intuitively, $N(w)$ is the set of propositions (set of states) that the agent believes at state w .

It is not hard to see that every Kripke frame $\langle W, R \rangle$ gives rise to a neighborhood frame. Define a neighborhood function N_R as follows: for $w \in W$, let $N_R(w) = \{X \mid \forall v, wRv \text{ implies } v \in X\}$. Now, given an arbitrary relation R , N_R is a **filter**. That is, for each $w \in W$, $N_R(w)$ is closed under supersets, finite

intersections and contains W . Furthermore if R is serial (for each w there is v such that wRv), then $\emptyset \notin N_R(w)$ (i.e., N_R is a proper filter). Finally, for each $w \in W$, $\cap N_R(w)$ is an element of $N_R(w)$. Conversely, any neighborhood frame N that is a proper filter such that for each $w \in W$, $\cap N(w) \in N(w)$, gives rise to a Kripke structures as follows: for each $w, v \in W$, say $wR_N v$ provided $v \in \cap N(w)$.

Our goal in this section is to formalize the Brandenburger-Keisler paradox. The strategy is to define a belief model intended to represent the beliefs of each agent about the other agent's beliefs (which are in turn beliefs about the first agent). Statements about the agents beliefs will written in a modal language and interpreted in these models. Using this machinery we can show that there is a formula that cannot be satisfied at any state. Our models will be two-sorted neighborhood models.

Definition 2 *An interactive neighborhood belief frame is a two-sorted structure $\mathcal{M} = \langle W, N^a, N^b \rangle$ where $W = W^a \cup W^b$, $N^a : W^a \rightarrow 2^{2^{W^b}}$ and $N^b : W^b \rightarrow 2^{2^{W^a}}$ where*

- *Each neighborhood function is a filter: Let $i \in \{a, b\}$, (i) for each $w \in W$, $X, Y \in N^i(w)$ implies $X \cap Y \in N^i(w)$; (ii) for each $w \in W$, $X \in N^i(w)$ and $X \subseteq Y$ implies $Y \in N^i(w)$; and (iii) for each $w \in W$, $W^i \in N^i(w)$.*
- *Each neighborhood function contains its core: Let $i \in \{a, b\}$, for each $w \in W$, $\cap N^i(w) \in N^i(w)$*
- *For each $x \in W^a$, $\emptyset \notin N^a(x)$*
- *For each $y \in W^b$, $\emptyset \notin N^b(y)$.*
- *There is $x \in W^a$ such that $\cap N^a(x) \neq W^b$*
- *There is $y \in W^b$ such that $\cap N^b(y) \neq W^a$*

Given an Ann state $x \in W^a$, the set $N^a(x)$ is the set of all of Ann's beliefs (about Bob) in state x , and vice versa (given a Bob state y , $N^b(y)$ is the set of Bob's beliefs about Ann). Thus $\cap N^a(x)$ is Ann's *strongest belief*, or assumption, at state x , and $\cap N^b(y)$ is Bob's *strongest belief*, or assumption, at state y . Of course, given the discussion above and the assumptions we are making about the neighborhoods, every interactive neighborhood frame is equivalent to a two-sorted Kripke structure. However, we prefer to work with neighborhood models since they are relevant for the discussion of axiomatization and make clear the assumptions we are making about the agents epistemic state. To that end, we

discuss each of the above assumptions. Much of the following discussion has been widely discussed in the literature on epistemic logic and so our discussion will be brief. See [9] or [?] for a more indepth discussion. First and foremost, the object of the agent's beliefs are propositions about the other agent's possible states of belief, i.e., the object of Ann's beliefs are sets of Bob states (similarly for Bob). For this reason, in the discussion below we identify knowing some formula ϕ with knowing its truth set (set of states where ϕ is true).

Logical Omniscience: The assumption that each agent's neighborhood function is a *proper* filter amounts to assuming the agents are logically omniscient in the following sense. First, the agents do not know any inconsistent facts. This follows since we assume that the emptyset is not in the agent's neighborhood and the neighborhood is closed under intersection. Thus if an agent knows ϕ and the agent knows ψ , then ϕ and ψ must be consistent, since the agent knows $\phi \wedge \psi$ and so their intersection cannot be empty. Second, the agents believe all of the logical consequences of its current state of belief. This following from the fact that the neighborhoods are closed under supersets. Finally, since for each $i \in \{a, b\}$ and $w \in W$, $W^i \in N^i(w)$, the agents believe all tautologies.

Informative: The last two properties amount to assuming that each agent has some nontrivial information about the other agent's beliefs. That is, if $\cap N^a(x) = W^b$ for all $x \in W^a$, then all of Ann's beliefs must be tautologies.

Let At be a set of propositional variables. Our language has the following syntactic form

$$p \mid \neg\phi \mid \phi \wedge \psi \mid \Box_i\phi \mid \boxplus_i\phi$$

where $p \in \text{At}$ and $i \in \{a, b\}$. A valuation is a function $V : \text{At} \rightarrow 2^W$ and an **interactive neighborhood belief model** is a tuple $\mathcal{M} = \langle W, N^a, N^b, V \rangle$ where V is a valuation and $\langle W, N^a, N^b \rangle$ is an interactive belief frame. Truth is defined as follows: for $x \in W^a$, $y \in W^b$ and $w \in W^a \cup W^b$ we have

- $\mathcal{M}, w \models p$ iff $w \in V(p)$
- $\mathcal{M}, w \models \neg\phi$ iff $\mathcal{M}, w \not\models \phi$
- $\mathcal{M}, w \models \phi \wedge \psi$ iff $\mathcal{M}, w \models \phi$ and $\mathcal{M}, w \models \psi$
- $\mathcal{M}, x \models \Box_a\phi$ iff $(\phi)^\mathcal{M} \in N^a(x)$.
- $\mathcal{M}, y \models \Box_b\phi$ iff $(\phi)^\mathcal{M} \in N^b(y)$
- $\mathcal{M}, x \models \boxplus_a\phi$ iff $(\phi)^\mathcal{M} = \cap N^a(x)$
- $\mathcal{M}, y \models \boxplus_b\phi$ iff $(\phi)^\mathcal{M} = \cap N^b(y)$.

We may write $w \models \phi$ for $\mathcal{M}, w \models \phi$ if it is clear which model is under consideration. At this stage it is useful to introduce some special propositional variables. Let \mathbf{A} denote the propositions W^a , that is $x \models \mathbf{A}$ iff $x \in W^a$. Similarly let $\mathbf{B} := \neg \mathbf{A}$ denote Bob states. Given a Bob state y , the set $\cap N^b(y)$ is Bob's assumption. Given an Ann state x , if $x \notin \cap N^b(y)$ then Bob's assumption is not true of state x . That is, Bob's assumption about x is wrong. For a fixed x , consider the set of all Bob states in which Bob's assumption about x is wrong: $\{y \mid x \notin \cap N^b(y)\}$. Call this set P_x and let \mathbf{P}_x represent this proposition, i.e., $v(\mathbf{P}_x) = P_x$. Now the set

$$D = \{x \mid \{y \mid x \notin \cap N^b(y)\} \in N^a(x)\}$$

is the set of states where Ann believes that Bob's assumption about here is wrong. Let \mathbf{D} represent this proposition, i.e., $V(\mathbf{D}) = D$. Our first Lemma is that it is impossible for Ann to believe that Bob assumes \mathbf{D} .

Lemma 3 *Suppose that \mathcal{M} is an interactive neighborhood belief model. Then there is no state $x \in W^a$ such that*

$$x \models \square_a \boxplus_b \mathbf{D}$$

Proof Suppose towards contradiction that \mathcal{M} is a belief model with an Ann state x such that $x \models \square_a \boxplus_b \mathbf{D}$. Then

$$(*) \quad \{y \mid D = \cap N^b(y)\} \in N^a(x)$$

Since $\emptyset \notin N^a(x)$, there is some y such that $D = \cap N^b(y)$. We arrive at a contradiction by showing that the set D cannot exist:

Suppose $x \in D$. Then by the definition of D , $\{y \mid x \notin \cap N^b(y)\} \in N^a(x)$. By (*) and the fact that $N^a(x)$ is a filter $\{y \mid D = \cap N^b(y)\} \cap \{y \mid x \notin \cap N^b(y)\} \in N^a(x)$. Since $x \in D, \{y \mid D = \cap N^b(y)\} \cap \{y \mid x \notin \cap N^b(y)\} = \emptyset$. But this contradicts the fact that $\emptyset \notin N^a(x)$.

Suppose $x \notin D$. Then $\{y \mid x \notin \cap N^b(y)\} \notin N^a(x)$. However, $\{y \mid D = \cap N^b(y)\} \subseteq \{y \mid x \notin \cap N^b(y)\}$ and by (*) and the fact that $N^a(x)$ is a filter, $\{y \mid x \notin \cap N^b(y)\} \in N^a(x)$. Contradiction. ■

This Lemma is essentially a restatement of the paradox from the Introduction. The proof illustrates exactly which assumptions about the agents state of beliefs

are needed. In particular, it is clear that each of the properties of a proper filter are used and if any of them are dropped then the proof will not go through. Furthermore, it is clear that the assumption operator \boxplus_b is crucial. For suppose that $x \models \Box_a \Box_b D$. Then $\{y \mid \cap N^b(y) \subseteq D\} \in N^a(x)$. From this assumption we cannot derive a contradiction. In particular, $\{y \mid \cap N^b(y) \subseteq D\}$ and $\{y \mid x \notin N^b(y)\}$ need not be disjoint. However, assuming $x \notin D$ still leads to a contradiction. Thus

Observation 4 *Let \mathcal{M} be an interactive neighborhood belief model. For each $x \in W^a$, if $x \models \Box_a \Box_b D$ then $x \models D$.*

Proof This follows from the proof of Lemma 3. We need only note that if $x \notin D$ then $\{y \mid D \in N^b(y)\} \subseteq \{y \mid x \notin \cap N^b(y)\}$. ■

In fact, in [7], Brandenburger and Keisler show precisely which statements cannot be assumed by agents in an interactive belief frame. We now proceed to prove an analogous theorem in this setting. We first need a lemma.

Lemma 5 *Suppose that \mathcal{M} is an augmented neighborhood interactive belief model. Then if there exists a state $x_1 \in \mathbf{A}$ such that $x_1 \models \boxplus_a \mathbf{B}$ and $x_2 \in W^a$ such that*

$$x_2 \models \Box_a \Box_b \Box_a \boxplus_b \mathbf{A}$$

Then,

$$x_2 \models \Box P_{x_2}$$

where $V(P_{x_2}) = \{y \mid x_2 \notin \cap N^b(y)\}$.

Proof Suppose $x_1 \models \boxplus_a \mathbf{B}$ and

$$x_2 \models \Box_a \Box_b \Box_a \boxplus_b \mathbf{A}$$

But suppose that $x_2 \not\models \Box P_{x_2}$. Then $V(P_{x_2}) \notin N^a(x_2)$ and so $\cap N^a(x_2) \not\subseteq V(P_{x_2})$. Hence there is a y_0 such that

1. $y_0 \in \cap N^a(x_2)$, and
2. $x_2 \in \cap N^b(y_0)$

By assumption, $\{y \mid y \models \Box_b \Box_a \boxplus_b W^a\} \in N^a(x_2)$. Hence by 1., $y_0 \models \Box_b \Box_a \boxplus_b W^a$. That is $\{x \mid x \models \Box_a \boxplus_b W^a\} \in N^b(y_0)$. By 2., $x_2 \models \Box_a \boxplus_b W^a$. That is, $\{y \mid \cap N^b(y) = W^a\} \in N^b(x_2)$. In particular, by 1

$$(*) \quad \cap N^b(y_0) = W^a.$$

Claim For each $x \in W^a$, $\{y \mid \cap N^b(y) = W^a\} \in N^a(x)$. Otherwise there is a $x' \in W^a$ such that $\{y \mid \cap N^b(y) = W^a\} \notin N^a(x')$. By (*), $x' \in \cap N^b(y_0)$ and so, $x' \models \Box_a \boxplus_b \mathbf{A}$. Hence $\{y \mid \cap N^b(y) = W^a\} \in N^a(x')$. Contradiction. Thus the claim is proved.

From the claim, $\{y \mid \cap N^b(y) = W^a\} \in N^a(x_1)$. Since, $\cap N^a(x_1) = W^b$, for each y , $\cap N^b(y) = W^a$. This contradicts the assumption that there is an $y \in W^b$ such that $\cap N^b(y) \neq W^a$.

■

Using this Lemma we can be more specific about precisely which statements cannot be assumed in an interactive belief model. Give a interactive belief frame \mathcal{M} , \mathcal{M} is said to have a **hole** at a formula ϕ provided ϕ is satisfiable in \mathcal{M} but $\boxplus_i \phi$ is not satisfiable for some $i \in \{a, b\}$. The interactive neighborhood frame is said to have a **big hole** at a formula ϕ if ϕ is satisfiable in \mathcal{M} but $\Box_i \phi$ is not satisfiable for some $i \in \{a, b\}$.

Theorem 6 *Every interactive neighborhood model \mathcal{M} has either a hole at \mathbf{A} , $\neg \mathbf{A}$, or \mathbf{D} ; or a big hole at $\boxplus_b \mathbf{A}$, $\Box_a \boxplus_b \mathbf{A}$, $\Box_b \Box_a \boxplus_b \mathbf{A}$, or $\boxplus_b \mathbf{D}$.*

Proof Suppose the statement is false. Thus there are no holes nor big holes at any of the above formulas. Since both \mathbf{A} and $\mathbf{B} := \neg \mathbf{A}$ are obviously satisfiable and there are no holes at these formulas, there are $x_1 \in W^a$ and $y_1 \in W^b$ such that $x_1 \models \boxplus_a \mathbf{B}$ and $y_1 \models \boxplus_b \mathbf{A}$. Since $\boxplus_b \mathbf{A}$ is satisfiable, there is x_2 such that $x_2 \models \Box_a \boxplus_b \mathbf{A}$. Hence there is y_2 such that $y_2 \models \Box_b \Box_a \boxplus_b \mathbf{A}$. Hence there is x_3 such that $x_3 \models \Box_a \Box_b \Box_a \boxplus_b \mathbf{A}$. By Lemma 5, since $x_1 \models \boxplus_a \mathbf{B}$, $x_3 \models \mathbf{D}$. Hence there is a y_3 such that $y_3 \models \boxplus_b \mathbf{D}$. Since there is no big hole at $\boxplus_b \mathbf{D}$, there must be x_5 such that $x_5 \models \Box_a \boxplus_b \mathbf{D}$. But this contradicts Lemma 3.

■

3.1 Towards an Axiomatization

In the previous section, the operator \boxplus_i is intended to represent an agent's assumption. In [7], it is asked whether a sound and complete axiomatization exists.

The main difficulty is whether or not the assumption operator should be treated as a modal operator. The problem is that at any state there is exactly one proposition that is assumed. A number of axioms suggest themselves. For example, it is easy to see that $\boxplus_i\phi \rightarrow \Box_i\phi$ is valid in any interactive belief model ($i \in \{a, b\}$). See [7] for a list of other possible axioms.

A similar question has been studied in the literature by a number of authors. A modal operator called the **window modality** has been studied by (among others) Gargov, Passy and Tinchev [10] and Goranko [11]. The operator is essentially the converse of the standard modal operator. Let $\langle W, R, V \rangle$ be a Kripke structure (i.e., W is a set of states, $R \subseteq W \times W$, and V is a valuation function). Interpret formulas of the form $\Box\phi$ as follows

$$\mathcal{M}, w \models \Box\phi \text{ iff } \forall v \text{ if } \mathcal{M}, v \models \phi \text{ then } wRv$$

See [2] for a complete discussion of this modality and pointers to the relevant literature.

The strategy we intend to pursue in the final version of this paper is not to think of the assumption operator as a modality. For simplicity, we consider the single agent case. Let \mathbb{P} be a set of propositional letters, W a set of states and $f : W \rightarrow \mathbb{P}$ a function. Intuitively, $f(w) \in \mathbb{P}$ is a proposition representing the agent's assumption at state w . Let At be a set of atomic propositions distinct from \mathbb{P} . A valuation function is a function $V : (\mathbb{P} \cup \text{At}) \rightarrow 2^W$ and a model is a tuple $\mathcal{M} = \langle W, \mathbb{P}, f, V \rangle$. We extend the basic modal language with formulas of the form $\boxplus P$ where $P \in \mathbb{P}$. Truth of the modal formulas is defined as follows: for each $w \in W$

$$\begin{aligned} \mathcal{M}, w \models \boxplus P &\text{ iff } f(w) = P \\ \mathcal{M}, w \models \Box\phi &\text{ iff } V(f(w)) \subseteq (\phi)^{\mathcal{M}} \end{aligned}$$

4 Using Hybrid Logic

Hybrid logic extends basic modal logic with special propositional variables, called nominals, intended to be “names” of possible worlds. See [4] and references therein for more information. The goal of this section is to study the Brandenburger-Keisler paradox from the hybrid logic point of view.

Definition 7 *A tuple $\mathcal{F} = \langle W, R_a, R_b, A, B \rangle$ is called a **belief frame** provided W is a (non-empty) set of states, A and B are subsets of W such that $W = A \cup B$ and $A \cap B = \emptyset$, $R_a \subseteq A \times B$ and $R_b \subseteq B \times A$ are both serial relations on W*

Throughout this sect, we use the following syntactic convention. The variables $x, x_1, x_2, x', x'', \dots$ will denote Ann variables. So, the formula $\forall x\phi(x)$ be shorthand for $\forall x(x \in A \Rightarrow \phi(x))$. Let $y, y', y'', y_1, y_2, \dots$ be Bob variables. Finally, we use z, z', z_1, z_2, \dots of arbitrary elements of W (they may belong to either A or B).

The language used to formalize the argument will be the full hybrid language. Let **Prop** be a countable set of propositional variables containing at least two designated variable P_A and P_B . Let **Nom** be a set of nominals and **Var** a set of variables. A formula can have the following syntactic form

$$\phi ::= P \mid i \mid \neg\phi \mid \phi \wedge \psi \mid \langle a \rangle\phi \mid \langle b \rangle\phi \mid @_i\phi \mid \forall x\phi(x)$$

where $i \in \mathbf{Nom}$ and $P \in \mathbf{Prop}$. Let $\vee, \rightarrow, [a], [b]$ and $\exists x\phi(x)$ be defined as usual. Following the above convention, define the following formulas

$$\forall x\phi(x) \stackrel{def}{=} \forall x(P_A \rightarrow \phi(x))$$

and

$$\forall y\phi(y) \stackrel{def}{=} \forall y(P_B \rightarrow \phi(y))$$

A model based on a frame \mathcal{F} is a tuple $\mathcal{M} = \langle \mathcal{F}, V \rangle$, where $V : \mathbf{Prop} \cup \mathbf{Nom} \rightarrow 2^W$ such that for each $i \in \mathbf{Nom}$, $|V(i)| = 1$ and $V(P_A) = A$ and $V(P_B) = B$. A substitution is a function $\sigma : \mathbf{Var} \rightarrow W$. Truth is defined relative to a state $w \in W$ and a substitution σ . We say that σ' is an **x-variant** of σ if $\sigma'(y) = \sigma(y)$ for all $y \neq x$. Truth is defined as follows. Let \mathcal{M} be an arbitrary belief model, $w \in W$ and σ a substitution:

- $\mathcal{M}, w \models_\sigma P$ iff $w \in V(P)$
- $\mathcal{M}, w \models_\sigma i$ iff $w \in V(i)$
- $\mathcal{M}, w \models_\sigma \neg\phi$ iff $\mathcal{M}, w \not\models_\sigma \phi$
- $\mathcal{M}, w \models_\sigma \phi \wedge \psi$ iff $\mathcal{M}, w \models_\sigma \phi$ and $\mathcal{M}, w \models_\sigma \psi$
- $\mathcal{M}, w \models_\sigma @_i\phi$ iff $\mathcal{M}, v \models_\sigma \phi$ where $v \in V(i)$
- $\mathcal{M}, w \models_\sigma \langle c \rangle\phi$ iff there is a $v \in W$ such that $wR_c v$ and $\mathcal{M}, v \models_\sigma \phi$ ($c \in \{a, b\}$)
- $\mathcal{M}, w \models_\sigma \forall x\phi(x)$ iff for each x -variant σ' , $\mathcal{M}, w \models_{\sigma'} \phi(x)$

The so called **binding operator** will be relevant for our study. Define $\downarrow x.\phi(x)$ as follows:

$$\downarrow x.\phi(x) \stackrel{def}{=} \exists x(x \wedge \phi(x))$$

Since this operator will play an important role in the following discussion, we give the definition of truth:

$\mathcal{M}, w \models_{\sigma} \downarrow x.\phi(x)$ iff $\mathcal{M}, w \models \sigma' \phi(x)$ where σ' is an x -variant of σ with $\sigma'(x) = w$

It is not hard to show that **B** is definable by the following formulas:

1. (two-sortedness) $@_i P_A \leftrightarrow @_i \neg P_B$
2. (two-sortedness a) $@_i \langle a \rangle j \rightarrow (@_i P_A \wedge @_j P_B)$
3. (two-sortedness b) $@_i \langle b \rangle j \rightarrow (@_x P_A \wedge @_j P_B)$
4. (seriality a) $@_i \langle a \rangle j$
5. (seriality b) $@_i \langle b \rangle j$

As discussed above the are “believes” and “assumes”. Recall that we say that a state x **believes** $\phi(y)$ provided $\forall y(P^a(x, y) \text{ implies } \phi(y))$. Here $\phi(y)$ may have y as a free variable, and so $\phi(y)$ is a statement *about the set of Ann-accessible states*. The translation to hybrid logic is straightforward and easily checked:

$$x \text{ believes } \phi(y) \text{ iff } @_x[a]\phi(y)$$

However, some care must be taken when expressing the formula $\phi(y)$. For example, to express “(at state x), Ann believes that Bob considers all of Ann’s states possible”, we use the formula

$$@_x[a] \downarrow y.\forall x.(@_y \langle b \rangle x)$$

The \downarrow operator bounds y to the current state. So the literal translation of the above formula is *at state x , in all a accessible states, y , all of the x states are accessible*.

Moving on to the slightly more complicated notion of “assumes”. Recall that x is said to **assume** a formula $\phi(y)$ provided $\forall y P^a(x, y) \text{ implies } \phi(y)$. With the help of the binding operators found in hybrid modal logic, translating assumes is also straightforward:

$$x \text{ assumes } \phi(y) \text{ iff } @_x(\forall y(@_x \langle a \rangle y \leftrightarrow \phi(y)))$$

For example, to express “(at state x), Ann *assumes* that Bob considers all of Ann’s states possible” use the following formula:

$$\@_x \forall y. (\@_x \langle a \rangle y \leftrightarrow \forall z. (\@_y \langle b \rangle z))$$

here z is an Ann variable (like x).

The following two lemmas are hybrid versions of Lemma 3 and Lemma 5. The proofs can be found in the appendix.

Lemma 8 *In all belief frames, if*

1. $\@_{x_1} \forall y. \@_{x_1} \langle a \rangle y$; and
2. $\@_{x_2} [a][b][a] \downarrow y. (\forall x. (\@_y \langle b \rangle x))$

are valid, then so is

$$\@_{x_2} [a] \neg \langle b \rangle x_2$$

Lemma 9 *In any model based on a belief frame, for any state x_0 ,*

$$\@_{x_0} [a] \downarrow y. (\forall x. (\@_y \langle b \rangle x \leftrightarrow \@_x ([a] \neg \langle b \rangle x)))$$

is false

We say that a model based on a belief frame has a **hole** at $\phi(y)$ iff there $\phi(y)$ is satisfiable, but not assumed by any agent. In other words, the conjunction of

1. $\@_x \phi(y)$; and
2. $\forall x. \neg (\@_x \forall y. (\@_x \langle a \rangle y \leftrightarrow \phi(y)))$

is valid. The model has a big hole provided, $\@_x \phi(y) \wedge \forall x. (\neg \@_x [a] \phi(y))$ is valid. As in Theorem 1, these previous two lemmas can be used to show precisely which formulas are holes or big holes.

A Proofs

The proofs in this appendix use a tableaux for quantified hybrid logic from [3]. The reader is referred to [3] for details of the proof system.

Lemma 10 *In all belief frames, if*

1. $@_{x_1} \forall y. @_{x_1} \langle a \rangle y$; and
2. $@_{x_2} [a][b][a] \downarrow y. (\forall x. (@_y \langle b \rangle x))$

are valid, then so is

$$@_{x_2} [a] \neg \langle b \rangle x_2$$

	0a. $@_{x_1} \forall y. @_{x_1} \langle a \rangle y$	Assumption 1
	0b. $@_{x_2} [a][b][a] \downarrow y. (\forall x. (@_y \langle b \rangle x))$	Assumption 2
	0c. $\neg @_{x_2} [a] \neg \langle b \rangle x_2$	Assumption 3
	1. $@_i \exists y \exists x. \neg @_y \langle b \rangle x$	Axiom
	2. $@_i \exists x. \neg @_y \langle b \rangle x$	1, \exists -rule
	3. $@_i \neg @_y \langle b \rangle x_0$	2, \exists -rule
	4. $\neg @_i @_{y_0} \langle b \rangle x_0$	3, \neg -rule
	5. $\neg @_{y_0} \langle b \rangle x_0$	4, $\neg @$ -rule
	6. $@_{x_2} \langle a \rangle y_1$	0c, $\neg[a]$
	6'. $\neg @_{y_1} \neg \langle b \rangle x_2$	Ditto
	7. $@_{y_1} \langle b \rangle x_2$	6', $\neg \neg$
Proof	8. $@_{y_1} [b][a] \downarrow y. (\forall x. (@_y \langle b \rangle x))$	6, 0b, $[a]$
	9. $@_{x_2} [a] \downarrow y. (\forall x. (@_y \langle b \rangle x))$	7, 0b, $[b]$
	10. $@_{y_1} \downarrow y. (\forall x. (@_y \langle b \rangle x))$	6, 0b, $[a]$
	11. $@_{y_1} \forall x. (@_{y_1} \langle b \rangle x)$	10, \downarrow
	12. $@_{y_1} @_{y_1} \langle b \rangle x_1$	11, \forall
	13. $@_{y_1} \langle b \rangle x_1$	12, $@$
	14. $@_{x_1} \langle a \rangle y_0$	0a, \forall
	15. $@_{x_1} [a] \downarrow y. (\forall x. (@_y \langle b \rangle x))$	13, 8, $[b]$
	16. $@_{y_0} \downarrow y. (\forall x. (@_y \langle b \rangle x))$	15, 14, $[a]$
	17. $@_{y_0} \forall x. (@_{y_0} \langle b \rangle x)$	16, \downarrow
	18. $@_{y_0} @_{y_0} \langle b \rangle x_0$	17, \forall
	19. $@_{y_0} \langle b \rangle x_0$	18, $@$

Contradiction 19,5

■

Lemma 11 *In any model based on a belief frame, for any state x_0 ,*

$$\@_{x_0}[a] \downarrow y.(\forall x.(@_y\langle b\rangle x \leftrightarrow @_x([a]\neg\langle b\rangle x)))$$

is false

	0.	$\@_{x_0}[a] \downarrow y.(\forall x.(@_y\langle b\rangle x \leftrightarrow @_x([a]\neg\langle b\rangle x)))$	Assumption
	1.	$\@_{x_0}\langle a\rangle y_0$	Seriality Axiom
	2.	$\@_{y_0} \downarrow y.(\forall x.(@_y\langle b\rangle x \leftrightarrow @_x([a]\neg\langle b\rangle x)))$	0, 1, [a]
Proof	3.	$\@_{y_0}(\forall x.(@_{y_0}\langle b\rangle x \leftrightarrow @_x([a]\neg\langle b\rangle x)))$	2, \downarrow
	4.	$\@_{y_0}(@_{y_0}\langle b\rangle x_0 \leftrightarrow @_{x_0}([a]\neg\langle b\rangle x_0))$	3, \forall
	5.	$\@_{y_0}(@_{y_0}\langle b\rangle x_0 \rightarrow @_{x_0}([a]\neg\langle b\rangle x_0))$	4, \wedge
	6.	$\@_{y_0}(@_{x_0}([a]\neg\langle b\rangle x_0) \rightarrow @_{y_0}\langle b\rangle x_0)$	Ditto

Left (5)

- a5 $\neg\@_{y_0}(@_{y_0}\langle b\rangle x_0)$
- b5 $\neg\@_{y_0}\langle b\rangle x_0$

Right (5)

- 5a $\@_{y_0}\@_{x_0}[a]\neg\langle b\rangle x_0$
- 5b $\@_{x_0}[a]\neg\langle b\rangle x_0$ 5a, @
- 5c $\@_{y_0}\neg\langle b\rangle x_0$ 5b, 1, [a]
- 5d $\neg\@_{y_0}\langle b\rangle x_0$ 5c $\neg\@$

Since both branches end with $\neg\@_{y_0}\langle b\rangle x_0$, call this formula (*), we need only close one of the branches (making sure to only find contradictions with formulas on both branches). The branch immediately splits using line 6:

Left (6)

- | | | |
|-----|--|------------------|
| a6 | $\neg\@_{y_0}\@_{x_0}[a]\neg\langle b\rangle x_0$ | |
| b6 | $\neg\@_{x_0}[a]\neg\langle b\rangle x_0$ | a6, $\neg\@$ |
| c6 | $\@_{x_0}\langle a\rangle y_1$ | b6, $\neg[a]$ |
| c6' | $\neg\@_{y_1}\neg\langle b\rangle x_0$ | Ditto |
| d6 | $\@_{y_1}\langle b\rangle x_0$ | c6', $\neg\neg$ |
| e6. | $\@_{y_1} \downarrow y.(\forall x.(@_y\langle b\rangle x \leftrightarrow @_x([a]\neg\langle b\rangle x)))$ | c6, 1, [a] |
| f6. | $\@_{y_1}(\forall x.(@_{y_1}\langle b\rangle x \leftrightarrow @_x([a]\neg\langle b\rangle x)))$ | e6, \downarrow |
| f6. | $\@_{y_1}(@_{y_1}\langle b\rangle x_0 \leftrightarrow @_{x_0}([a]\neg\langle b\rangle x_0))$ | e6, \forall |
| g6. | $\@_{y_1}(@_{y_1}\langle b\rangle x_0 \rightarrow @_{x_0}([a]\neg\langle b\rangle x_0))$ | f6, \wedge |
| h6. | $\@_{y_1}(@_{x_0}([a]\neg\langle b\rangle x_0) \rightarrow @_{y_0}\langle b\rangle x_0)$ | Ditto |

Left (g6)
 $ag6 \quad \neg @_{y_1} (@_{y_1} \langle b \rangle x_0)$
 $bg6 \quad \neg @_{y_1} \langle b \rangle x_0$ $ag6, \neg @$
Contradiction d6, bg6

Right (g6)
 $g6a \quad @_{y_1} @_{x_0} ([a] \neg \langle b \rangle x_0)$
 $g6b \quad @_{x_0} [a] \neg \langle b \rangle x_0$ $g6a, @$
 $g6c \quad @_{y_1} \neg \langle b \rangle x_0$ $c6, g6b, [a]$
 $g6d \quad \neg @_{y_1} \langle b \rangle x_0$ $g6c, \neg @$
Contradiction g6d, d6

Right (6)
 $6a \quad @_{y_0} @_{y_0} \langle b \rangle x_0$
 $6b \quad @_{y_0} \langle b \rangle x_0$ $6a, @$
 $6c \quad$ **Contradiction 6b, ***

■

References

- [1] AUMANN, R. Interactive epistemology I: knowledge. *International Journal of Game Theory* 28 (1999), 263–300.
- [2] BLACKBURN, P., DE RIJKE, M., AND VENEMA, Y. *Modal Logic*. Cambridge University Press, 2002.
- [3] BLACKBURN, P., AND MARX, M. Tableaux for quantified hybrid logic. In *Automated Reasoning with Analytic Tableaux and Related Methods* (2002).
- [4] BLACKBURN, P., AND SELIGMAN, J. Hybrid languages. *Journal of Logic, Language and Information* 4 (1995), 251 – 272.
- [5] BONANNO, G., AND BATTIGALLI, P. Recent results on belief, knowledge and the epistemic foundations of game theory. *Research in Economics* 53, 2 (June 1999), 149–225.
- [6] BRANDENBURGER, A. The power of paradox: some recent developments in interactive epistemology. *International Journal of Game Theory* (2006 (forthcoming)).

- [7] BRANDENBURGER, A., AND KEISLER, H. An impossibility theorem on beliefs in games. forthcoming in *Studia Logica*, available at pages.stern.nyu.edu/~abranden/itbg072904.pdf, July 2004.
- [8] CHELLAS, B. *Modal Logic: An Introduction*. Cambridge University Press, Cambridge, 1980.
- [9] FAGIN, R., HALPERN, J., MOSES, Y., AND VARDI, M. *Reasoning about Knowledge*. The MIT Press, 1995.
- [10] GARGOV, G., PASSY, S., AND TINCHEV, T. Modal environment for boolean speculations. In *Mathematical Logic and its Applications*, D. Skořdev, Ed. Plenum Press, 1987.
- [11] GORANKO, V. Modal definability in enriched languages. *Notre Dame Journal of Formal Logic* 31 (1996).
- [12] HARSANYI, J. C. Games with incomplete information played by bayesian players parts I-III. *Management Sciences* 14 (1967).
- [13] M.O.L. BACHARACH, L. A. GERARD-VARET, P. M., AND SHIN, H. S., Eds. *Epistemic logic and the theory of games and decisions*. Theory and Decision Library, Kluwer Academic Publishers, 1997.
- [14] MONTAGUE, R. Universal grammar. *Theoria* 36 (1970), 373 – 398.
- [15] OSBORNE, M., AND RUBINSTEIN, A. *A Course in Game Theory*. The MIT Press, 1994.
- [16] SCOTT, D. Advice in modal logic. In *Philosophical Problems in Logic*. K. Lambert, 1970, pp. 143 – 173.
- [17] STALNAKER, R. Belief revision in games: forward and backward induction. *Mathematical Social Sciences*, 36 (1998), 31 – 56.