Economic Theory **19**, 145–156 (2002)

Economic Theory

A note on the wise girls puzzle*

Mariko Yasugi[1] **and Sobei H. Oda**[2]

[1] Faculty of Science, Kyoto Sangyo University, Kita-ku, Kyoto 603-8555, JAPAN
(e-mail: yasugi@cc.kyoto-su.ac.jp)
[2] Faculty of Economics, Kyoto Sangyo University, Kita-ku, Kyoto 603-8555, JAPAN
(e-mail: oda@cc.kyoto-su.ac.jp)

Received: August 1, 2000; revised version: May 15, 2001

Summary. This article analyzes the *two wise girls puzzle*, which is a simpler variant of the so-called *three wise men puzzle*, with some *proof-theoretic* tools. We formulate the puzzle in an epistemic logic. Our chief assumption is that the reasoning ability of each player of the puzzle is equivalent to what is described by the epistemic logic. We will interpret the behaviors of the players in the puzzle in terms of *unprovability* of certain statements. The proof-theoretic tools we employ are consequences of a *meta-theorem*, known as the *cut elimination theorem*.

Keywords and Phrases: Puzzle, Propositional calculus, Belief operator, Proof-theory, Cut elimination, Unprovability.

JEL Classification Numbers: C69, C79, D82.

1 Introduction

In this article we take up the *two wise girls puzzle*, which is a simpler variant of the so-called *three wise men puzzle*. Puzzles of this type have been analyzed in a number of references, to which we will add an interpretation in terms of some meta-theorems on a logical system.

We formulate the puzzle as follows.

* The authors are indebted to Osamu Takaki for his careful reading of the original manuscript, giving us some useful comments. We are also grateful to H. Ono, M. Sato, H. Hisamoto, J. Iritani and T. Matsuhisa for their interests and discussions. K. Matsumoto and M. Takano kindly provided us with some chronological information on proof-theoretic results in modal logics. Our special thanks are due to M. Kaneko, with whose advice we have been able to improve this article.
Correspondence to: M. Yasugi

Two girls are seated, facing the same direction, so that the first girl is seated behind the second girl; the girls are put on a white hat on their heads; the first girl can see the second girl's hat but not conversely, and neither can see her own hat (see Fig. 1). The girls are told by the observer that at least one of them wears a white hat. The first girl is asked by the observer: "Do you know if your hat is white?" She answers "No! I do not know." Then the second girl is asked the same question, and she answers "Yes, I know."

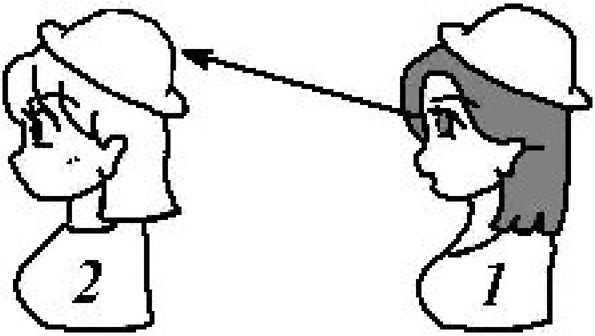

Figure 1 (illustrated by T. Kadota)

In this article, we will analyze the process of the puzzle in terms of *proof-theory* of a logical system which represents the reasoning abilities of the girls in the puzzle.

There are two points at issue, which are inherent in the puzzles of this type. They are more tangible in our simpler version.

The first point is that a certain ambiguity is hidden in the answer of the first girl, which is not discussed in the usual treatment of the puzzle. Namely, the first girl in fact cannot reach the conclusion "no" with her reasoning, and hence she either remains silent or gives up reasoning and answers "No!"

The second point is that the second girl must interpret the first girl's reaction in order to reach the right conclusion. In other words, she cannot deduce the correct answer without receiving the information from the behavior of the first girl.

We will give our version of the inability of the girls as explained above in terms of *unprovability* of certain statements.

Including this introduction, this article is composed of five sections. In Section 2, we will discuss our observation of the puzzle in more detail.

In Section 3, we present the logical system within which the girls can reason. The system, which is formulated in the *sequential calculus*, corresponds to the modal logic $\mathrm{KD4}^2$. Proof-theoretic tools for our purpose are also explained. A most useful tool is the meta-theorem called the *cut elimination theorem*. (The *cut elimination theorem* was originally proved for a system of classical predicate logic by Gentzen [2] and has been regarded as the central tool in proof-theory. It is known that the cut-elimination theorem holds for $\mathrm{KD4}^2$, and we will use this fact in some steps of our subsequent discussion.) Several consequences of the theorem are also listed here and in Section 4.

In Section 4, we interpret two girls' behaviors in terms of the proof-theoretic tools presented in Section 3.

In Section 5, we mention some remarks: the notion of *not knowing* (or, not believing); an *alternative interpretation* of the puzzle; the *solvability* of the puzzle; the *general case* of *n* girls.

The only background assumed in this article is some basic knowledge in the classical propositional calculus. Some texts are listed as references.

2 Discussion on two wise girls puzzle

We will henceforth call the first girl *Player 1* and the second girl *Player 2* .

Let us discuss the points at issue in the puzzle as were explained in Introduction.

First, *Player 1* would try to derive, from the pieces of information she has, the conclusion that she knows her hat is white by using logical reasoning, unsuccessfully. On the other hand, she cannot derive the negative conclusion either. In these circumstances, she might remain silent, or else, after some long search of a correct answer, she might give it up and answer "No!"

Next, *Player 2* might try to derive an answer from her initial information, unsuccessfully. Then she would try to interpret *Player 1* 's reaction in order to reach the right conclusion. Whether *Player 1* answers "No!" or remains silent, a reasonable player in the position of *Player 2* would interpret it as "*Player 1* does not know that she wears a white hat." With this interpretation, *Player 2* can conclude that she knows she wears a white hat.

The last process of the puzzle, that is, *Player 2* derives her conclusion, can be demonstrated straightforward. What must be treated with care is the inability of a player to derive an answer. That is, one needs some device in order to show that *Player 1* cannot derive any conclusion and that *Player 2* cannot derive a conclusion without waiting for *Player 1* 's reaction.

The objective of this article is to analyze the *two wise girls puzzle* by formulating it accurately in a formal language and by adopting *proof-theoretic tools*. For this purpose, we first formulate the statements of the puzzle in a language of the propositional calculus with the *belief operators*, which express that one *believes* (knows) a fact. In order to analyze the process of the puzzle, we *assume* that the players can reason logically. We will see that *Player 1* cannot logically derive (from the pieces of information she has) that she knows she wears a white hat, by showing the logical *unprovability* of the statement.

Then *Player 2* 's interpretation that *Player 1* does not know if her hat is white can be justified. *Player 2* would then proceed to logically derive that her own hat is white.

In the subsequent sections, we will formulate a logical system in which the players reason, and theoretically justify the players' reactions.

Before getting into an exact treatment, let us note the following.

In the standard "three wise men puzzle" (or muddy children puzzle), the men look at each other and answer simultaneously. See, for example, Sato [14]. We have taken our simpler version, since our objective is not to solve the puzzle, but to examine the solving process. With a simpler version, the essence of the solving process can be distilled.

3 Logical system and tools

We will set a language $\mathscr{L}$ with which one can express the statements of the puzzle, and then a logical (reasoning) system in which *Player 2* 's solution can be deduced. We adopt a formulation of the system by Kaneko [8], Section 4.4.

Except for propositional connectives, we need an operator B_i (called a belief operator) for each player, $i = 1, 2$. The logical system to be defined is KD4^2, a system of modal logic.

Although the system is described in Section 4.4 of Kaneko [8], we will present the definitions for the reader's convenience. For the modal logic, one can also refer to Chellas [1], Gerbrandy [3], Halpern and Moses [4], Hughes and Cresswell [7], Kaneko and Nagashima [9], Ohnishi and Matsumoto [13] and Sato [14]. For basic background in logical systems, we list Gentzen [2], Hayashi [5] and Kleene [10].

Definition 1 (Language and system)

1) The language $\mathscr{L}$: We prepare two propositional symbols $iW, i = 1, 2$. iW is read as "*Player i* wears a white hat."

The propositional connectives are $\neg$ (not), $\wedge$ (and), $\vee$ (or), and $\Rightarrow$ (implies). Parentheses "(" and ")" are also assumed.

Belief operators B_1 ("*Player 1* believes that") and B_2 ("*Player 2* believes that") are added.

2) Formulas of $\mathscr{L}$: $\mathscr{L}$*-formulas* are defined as follows.

2.1) Propositional symbols $1W$ and $2W$ are (atomic) $\mathscr{L}$-formulas.

2.2) If A is an $\mathscr{L}$-formula, then so are $(\neg A)$ and $\mathrm{B}_i(A), i = 1, 2$.

2.3) If A and B are $\mathscr{L}$-formulas, then so are $(A \wedge B), (A \vee B)$ and $(A \Rightarrow B)$.

Parentheses may be abbreviated when confusion is not likely, e.g. $\mathrm{B}_i A$ and $A \wedge B$.

3) Γ, $\Delta, \cdots$ each denotes a finite set of formulas (possibly empty). A set such as $\{A_1, A_2, \cdots, A_k\}$ will be sometimes written as $A_1, A_2, \cdots, A_k$, and the union of sets such as $\Gamma \cup \Delta$ may be written as Γ, Δ. For example, we may write Γ, A instead of $\Gamma \cup \{A\}$. We will also write $\mathrm{B}_i \Gamma$ for the set $\{\mathrm{B}_i A : A \in \Gamma\}$.

Using the notation above, we will introduce an expression called a *sequent*. For that purpose, we introduce a new symbol $\rightarrow$.

An expression of the form $\Gamma \rightarrow \Delta$ is called a *sequent*. Γ and Δ are respectively called the *antecedent* and the *succedent* of the sequent. A sequent $F_1, F_2, \cdots, F_m \rightarrow G_1, G_2, \cdots, G_n$ is intended to have the same meaning as the formula $F_1 \wedge F_2 \wedge \cdots \wedge F_m \Rightarrow G_1 \vee G_2 \vee \cdots \vee G_n$. The notion of a sequent is introduced for technical usefulness. We may write $\Gamma, F \rightarrow \Theta, G$ instead of $\Gamma \cup \{F\} \rightarrow \Theta \cup \{G\}$.

4) Initial sequents: A sequent of the form $A \rightarrow A$ for any formula A is called an *initial sequent*. (Such a sequent represents a tautology $A \Rightarrow A$.)

5) The logical inferences can be classified into three categories. The first one is called *thin*, which infers a *thinned* sequent from a given sequent. The second one is called *cut*, which cuts out a formula common to the succedent of a sequent and the antecedent of another. The third one consists of *propositional* inferences, which introduce propositional connectives in the antecedents or in the succedents.

$$\frac{\Gamma \rightarrow \Theta}{\Delta, \Gamma \rightarrow \Theta, \Lambda}\ (\mathit{thin}) \qquad \frac{\Delta \rightarrow \Theta, A \quad A, \Gamma \rightarrow \Lambda}{\Delta, \Gamma \rightarrow \Theta, \Lambda}\ (\mathit{cut})$$

$$\frac{A, \Gamma \rightarrow \Theta}{A \wedge B, \Gamma \rightarrow \Theta}\ (\wedge \rightarrow)_l \qquad \frac{B, \Gamma \rightarrow \Theta}{A \wedge B, \Gamma \rightarrow \Theta}\ (\wedge \rightarrow)_r$$

$$\frac{\Gamma \rightarrow \Theta, A \quad \Gamma \rightarrow \Theta, B}{\Gamma \rightarrow \Theta, A \wedge B}\ (\rightarrow \wedge)$$

$$\frac{A, \Gamma \rightarrow \Theta \quad B, \Gamma \rightarrow \Theta}{A \vee B, \Gamma \rightarrow \Theta}\ (\vee \rightarrow)$$

$$\frac{\Gamma \rightarrow \Theta, A}{\Gamma \rightarrow \Theta, A \vee B}\ (\rightarrow \vee)_l \qquad \frac{\Gamma \rightarrow \Theta, B}{\Gamma \rightarrow \Theta, A \vee B}\ (\rightarrow \vee)_r$$

$$\frac{\Gamma \rightarrow \Theta, A \quad B, \Gamma \rightarrow \Theta}{A \Rightarrow B, \Gamma \rightarrow \Theta}\ (\Rightarrow \rightarrow) \qquad \frac{A, \Gamma \rightarrow \Theta, B}{\Gamma \rightarrow \Theta, A \Rightarrow B}\ (\rightarrow \Rightarrow)$$

$$\frac{\Gamma \rightarrow \Theta, A}{\neg A, \Gamma \rightarrow \Theta}\ (\neg \rightarrow) \qquad \frac{A, \Gamma \rightarrow \Theta}{\Gamma \rightarrow \Theta, \neg A}\ (\rightarrow \neg)$$

6) Belief Inference: The belief operator is introduced by the following rule.

$$\frac{\Gamma, \mathrm{B}_i(\Delta) \rightarrow \Theta}{\mathrm{B}_i(\Gamma \cup \Delta) \rightarrow \mathrm{B}_i(\Theta)}\ (\mathrm{B}_i \rightarrow \mathrm{B}_i)$$

Here $i = 1, 2$, and Θ has at most one formula.

7) Upper and lower sequents: A sequent above the line of an inference is called an (the) *upper sequent* of the inference, and the one below the line is called the *lower sequent* of the inference. An upper sequent is an assumption and a lower sequent is the conclusion of an inference.

8) Proof-figure: A proof in $\mathrm{KD4}^2$ is a tree with a sequent at each node, where any topmost sequent is an initial sequent (cf. 4) above) and the sequents on the nodes are connected by the inferences in 5) above. A proof in this context is usually called a *proof-figure*. Some examples of proof-figures will be given in Sections 3 and 4.

9) Provability: A sequent $\Gamma \to \Delta$ is said to be *provable* in the system $\mathrm{KD4}^2$ if there is a proof-figure whose lowest sequent is $\Gamma \to \Delta$. It is said to be *unprovable* if it is not provable. A formula A is said to be provable if the sequent $\to A$ is provable.

The fact that "$\Gamma \to \Delta$ is provable in $\mathrm{KD4}^2$" is denoted by $\vdash\ \Gamma \to \Delta$, and the fact that "it is not provable" is denoted by $\nvdash \Gamma \to \Delta$.

Remark 1) Provability and unprovability as defined above are meta-notions.
2) A finite set of formulas Γ is *inconsistent* in $\mathrm{KD4}^2$ if $\Gamma \to$ is provable. Γ is said to be *consistent* otherwise. (Logically, $\Gamma \to$ and $\Gamma \to A\wedge\neg A$ are equivalent, and hence $\Gamma \to$ expresses that the formulas in Γ lead to a contradiction. The consistency of Γ can therefore be expressed as $\nvdash \Gamma \to$.)
3) A study of meta-notions is called *meta-mathematics*, and, if the method of such a study is a syntactic one, that is, without referring to truth values of formulas, it is called *proof-theory*. A meta-notion which is shown to hold by meta-mathematics is called a *meta-theorem*.

The fundamental meta-theorem for the system $\mathrm{KD4}^2$ is the following.

Theorem 1 (The cut elimination theorem for KD4^2) If $\vdash\ \Gamma \to \Delta$, then there is a *cut-free* proof-figure of $\Gamma \to \Delta$.

The present article can be read without knowing the proof of the theorem, and hence we do not include the proof here. It is included in Kaneko and Nagashima [9], the finitary part of whose treatment corresponds to $\mathrm{KD4}^2$. We refer the interested reader to Sections 11 and 12 in Yasugi and Oda [15], where a detailed cut elimination proof for $\mathrm{KD4}^2$ is presented.

It is an immediate consequence of Theorem 1 that $\mathrm{KD4}^2$ is *consistent*. Namely, no contradiction is provable in $\mathrm{KD4}^2$. (A contradiction is a formula of the form $A \wedge \neg A$ and the system can be said to be consistent if a sequent of the form $\to A \wedge \neg A$ is *not* provable.)

Lemmas 1$\sim$3 are crucial in our analysis. Lemma 2 is a consequence of the cut-elimination theorem.

Lemma 1 (Elimination lemma: cf. Kaneko and Nagashima [9]) Let $\epsilon_i\Gamma$ denote the result of eliminating all the occurrences of B_i (as well as superfluous parentheses) from Γ, and let $\epsilon\Gamma$ denote the result of eliminating all the belief operators from Γ.

If a sequent $\Gamma \to \Theta$ is provable in KD4^2, then $\epsilon_i\Gamma \to \epsilon_i\Theta$ is provable in KD4^2 without an application of ($\mathrm{B}_i \to \mathrm{B}_i$), and also $\epsilon\Gamma \to \epsilon\Theta$ is provable in the classical propositional logic.

Definition 2 (Separation: cf. Kaneko and Nagashima [9]) A formula A is called B_i*-atomic* if the outermost symbol of A is B_i. A formula A is called a B_i-formula if it is constructed from B_i-atomic formulas by applications of propositional connectives. A is called a B_{-i}-formula if B_i occurs in A only in the scope of a B_j for $j \neq i$.

For example, $\mathrm{B}_1(\mathrm{B}_2 B) \wedge \mathrm{B}_1(B \Rightarrow \mathrm{B}_1 C)$ is a B_1-formula, where $\mathrm{B}_1(\mathrm{B}_2 B)$ and $\mathrm{B}_1(B \Rightarrow \mathrm{B}_1 C)$ are B_1-atomic subformulas. $\mathrm{B}_2(B \vee \mathrm{B}_1 C) \Rightarrow 1W$ is a B_{-1}-formula. $\mathrm{B}_1 B \wedge \mathrm{B}_2 C$ is neither a B_1-formula nor a B_{-1}-formula.

A sequent is called B_i*-separable* if it consists of B_i-formulas and B_{-i}-formulas.

Lemma 2 (Separation lemma: cf. Theorem 3.3 of Kaneko and Nagashima [9]) Let $\Gamma, \Delta \to \Theta, \Lambda$ be a B_i-separable sequent, where Γ and Θ each consists of B_i-formulas and Δ and Λ each consists of B_{-i}-formulas. If $\Gamma, \Delta \to \Theta, \Lambda$ is provable in KD4^2, then $\Gamma \to \Theta$ or $\Delta \to \Lambda$ is provable in KD4^2.

For the proof, apply the cut-elimination theorem to any proof-figure. Then simply check the claimed fact for each sequent in a cut-free proof-figure downward, starting with initial sequents. The reader who is interested in a detailed proof is invited to look at Section 2 of Yasugi and Oda [15].

The following is a well-known fact in the classical propositional calculus, which also holds for the present system KD4^2.

Lemma 3 (Implication distribution lemma) Suppose

$$\vdash\ A_1 \Rightarrow B_1, A_2 \Rightarrow B_2, \Gamma \to \Delta.$$

Then all the following four sequents are provable: (1) $\Gamma \to \Delta, A_1, A_2$; (2) $B_1, B_2, \Gamma \to \Delta$; (3) $B_1, \Gamma \to \Delta, A_2$; (4) $B_2, \Gamma \to \Delta, A_1$.

Note. This property holds for an arbitrary number of pairs $\{A_j, B_j\}_{j=1,2,\cdots,n}$.

As an example, we present a proof-figure of (1) from $\vdash\ A_1 \Rightarrow B_1, A_2 \Rightarrow B_2, \Gamma \to \Delta$.

$$\dfrac{\dfrac{\dfrac{A_2 \to A_2}{A_2 \to A_2, B_2}\,(\mathit{thin})}{\to A_2, A_2 \Rightarrow B_2}\,(\to\Rightarrow) \qquad \dfrac{\dfrac{\dfrac{A_1 \to A_1}{A_1 \to A_1, B_1}\,(\mathit{thin})}{\to A_1, A_1 \Rightarrow B_1}\,(\to\Rightarrow) \quad A_1 \Rightarrow B_1, A_2 \Rightarrow B_2, \Gamma \to \Delta}{A_2 \Rightarrow B_2, \Gamma \to \Delta, A_1}\,(\mathit{cut})}{\Gamma \to \Delta, A_1, A_2}\,(\mathit{cut})$$

Lemma 4 (Cut within belief)

$$\mathrm{B}_i(A \Rightarrow B), \mathrm{B}_i(B \wedge C \Rightarrow D) \to \mathrm{B}_i(A \wedge C \Rightarrow D)$$

is provable in KD4^2. (A sequent without $\wedge C$ is also provable.)

This is an immediate consequence of

$$A \Rightarrow B, B \wedge C \Rightarrow D \to A \wedge C \Rightarrow D$$

by an application of ($\mathrm{B}_i \to \mathrm{B}_i$).

4 Reasoning process of players

Let us express more precisely what was discussed in Introduction and Section 2. We will show that *Player 1* cannot obtain the exact answer by reasoning in KD4^2. We will adopt the proof-theoretic tools in Section 3 to our arguments.

We will write the fact that *Player 1* does not believe a fact A as $\neg\mathrm{B}_1 A$. Then $\mathrm{B}_2\neg\mathrm{B}_1 1W$ expresses the fact that "*Player 2* believes that *Player 1* does not believe she (*Player 1*) wears a white hat."

Let W_0 denote $1W \vee 2W$. ("*Player 1* or *Player 2* wears white.") Then, the initial *belief set* for *Player 1* , Γ_1, is expressed as

$$\Gamma_1 = \{\mathrm{B}_1 W_0, \mathrm{B}_1 2W\}$$

Notice that Γ_1 consists of B_1-formulas (cf. Definition 2).

Player 2 's initial belief set, Γ_2, is expressed as follows.

$$\Gamma_2 = \{\mathrm{B}_2 W_0, \mathrm{B}_2(\mathrm{B}_1 W_0), \mathrm{B}_2(2W \Rightarrow \mathrm{B}_1 2W), \mathrm{B}_2(\neg 2W \Rightarrow \mathrm{B}_1\neg 2W)\}$$

Notice that Γ_2 consists of B_2-formulas.

We can show the consistency of each player's initial belief set, that is, (i) $\nvdash \Gamma_1 \to$ and (ii) $\nvdash \Gamma_2 \to$. These two facts can be established similarly to the proposition below.

First, we prove the following proposition, which states that *Player 1* can reach no definite answer.

Proposition 1 (1) $\nvdash \Gamma_1 \to \mathrm{B}_1 1W$

(2) $\nvdash \Gamma_1 \to \mathrm{B}_1\neg 1W$

Proof. (1) Suppose $\Gamma_1 \to \mathrm{B}_1 1W$ were provable. Then, by Lemma 1 (B_1-elimination), $W_0, 2W \to 1W$ would be provable in the classical propositional calculus, which is easily shown to be impossible. (The unprovability of $W_0, 2W \to 1W$ in the classical propositional calculus can be shown by constructing a counter-model of it. See Kaneko [8], Section 3. It can also be shown by examining cut-free proof-figures.)

(2) can be shown similarly.

Interpretation of *Player 1*'s behavior Due to (1) of Proposition 1, *Player 1* can never reach the conclusion $\mathrm{B}_1 1W$ from her belief set Γ_1 with her logical ability. So, if she tried, she would keep searching for a proof in vain. Such a state of affairs would make *Player 1* remain silent. However, since she is asked a question and since she would get tired of her search, it would be natural to assume that *Player 1* eventually gives up thinking and answers that she does not know if she wears white.

The fact that *Player 1* cannot logically reach $\Gamma_1 \to \mathrm{B}_1 1W$ differs from that she can reach the conclusion $\Gamma_1 \to \mathrm{B}_1 \neg 1W$. In fact, (2) of Proposition 1 states that she *cannot*.

Now, we go on to *Player 2* 's problem. Without taking *Player 1* 's answer into account, *Player 2* cannot reach any definite answer.

Proposition 2 (1) $\nvdash \Gamma_2 \to \mathrm{B}_2 2W$
(2) $\nvdash \Gamma_2 \to \mathrm{B}_2 \neg \mathrm{B}_1 1W$
(3) $\nvdash \Gamma_2, \mathrm{B}_2 \neg \mathrm{B}_1 1W \to$

Proof. (2) Suppose $\Gamma_2 \to \mathrm{B}_2 \neg \mathrm{B}_1 1W$ were provable. First apply Lemma 1 with $i = 2$ (B_2-elimination) to obtain

$$W_0, \mathrm{B}_1 W_0, 2W \Rightarrow \mathrm{B}_1 2W, \neg 2W \Rightarrow \mathrm{B}_1 \neg 2W \to \neg \mathrm{B}_1 1W$$

This is equivalent to

$$W_0, \mathrm{B}_1 W_0, 2W \Rightarrow \mathrm{B}_1 2W, \neg 2W \Rightarrow \mathrm{B}_1 \neg 2W, \mathrm{B}_1 1W \to$$

By Lemma 3 (Implication distribution lemma) applied to $2W \Rightarrow \mathrm{B}_1 2W$ and $\neg 2W \Rightarrow \mathrm{B}_1 \neg 2W$, we obtain that, in particular,

$$W_0, \mathrm{B}_1 W_0, \mathrm{B}_1 2W, \mathrm{B}_1 1W \to \neg 2W$$

must be provable. By Lemma 2 (Separation) with respect to B_1-formulas, either $W_0 \to \neg 2W$ or $\mathrm{B}_1 W_0, \mathrm{B}_1 2W, \mathrm{B}_1 1W \to$ must be provable. The first one is impossible. As for the second one, applying Lemma 1 with $i = 1$ (B_1-elimination), $W_0, 2W, 1W \to$ must be provable, but this is impossible. (These impossibilities are also shown by constructing counter-models.)

(1) and (3) can be proved similarly.

Interpretation of *Player 2*'s behavior Due to (1) of Proposition 2, *Player 2* cannot logically derive the conclusion that she wears a white hat from her initial belief set. Upon hearing *Player 1* 's answer or interpreting her silence as a negative answer, *Player 2* interprets it as $\neg \mathrm{B}_1 1W$. The fact that *Player 2 believes* $\neg \mathrm{B}_1 1W$ can be expressed as $\mathrm{B}_2 \neg \mathrm{B}_1 1W$. *Player 2* expands her belief set by adding the new piece of belief. The fact that *Player 2* indeed needs this new piece of belief, that is, that she had to wait for *Player 1* 's reaction in order to assume $\mathrm{B}_2 \neg \mathrm{B}_1 1W$, can be assured with (2) of Proposition 2. The fact that *Player 2* can add $\mathrm{B}_2\neg\mathrm{B}_1 1W$ to her belief set without causing a contradiction is assured with (3) of Proposition 2.

Now, from the expanded belief set, she would reason with her logical ability to obtain her conclusion.

Expanded belief set and conclusion Let us denote *Player 2* 's new belief set by $\Gamma_2' = \Gamma_2 \cup \{\mathrm{B}_2\neg\mathrm{B}_1 1W\}$. By (3) of Proposition 2, Γ_2' is consistent. Using Γ_2', *Player 2* will now deduce her conclusion. We will present *Player 2* 's reasoning step by step in order to show that the system KD4^2 is adequate for that purpose.

Proposition 3 (Conclusion)

$$\vdash \quad \Gamma_2' \to \mathrm{B}_2 2W$$

Proof. Let F denote one of the formulas in 1∼8 below. We show that $\Gamma_2' \to F$ can be derived succesively, so that the desired formula $\mathrm{B}_2 2W$ will be reached at the end (in 8 below). ($\bot$ will denote a contradiction, that is, any formula of the form $X \wedge \neg X$.)

1. $\mathrm{B}_2(\neg 2W \Rightarrow \mathrm{B}_1\neg 2W)$
2. $\mathrm{B}_2(\mathrm{B}_1\neg 2W \Rightarrow \mathrm{B}_1 1W)$
3. $\mathrm{B}_2(\mathrm{B}_1 1W \wedge \neg\mathrm{B}_1 1W \Rightarrow \bot)$
4. $\mathrm{B}_2(\neg 2W \wedge \neg\mathrm{B}_1 1W \Rightarrow \bot)$
5. $\mathrm{B}_2(\neg\mathrm{B}_1 1W \Rightarrow \neg\neg 2W)$
6. $\mathrm{B}_2(\neg\neg 2W \Rightarrow 2W)$
7. $\mathrm{B}_2(\neg\mathrm{B}_1 1W \Rightarrow 2W)$
8. $\mathrm{B}_2 2W$

1 is an assumption in Γ_2'. 2 is obtained from $W_0, \neg 2W \to 1W$ by applications of ($\mathrm{B}_1 \to \mathrm{B}_1$) and ($\mathrm{B}_2 \to \mathrm{B}_2$). (Recall that W_0 denotes $1W \vee 2W$.) 3 is obtained from $\to (\mathrm{B}_1 1W \wedge \neg\mathrm{B}_1 1W \Rightarrow \bot)$ by ($\mathrm{B}_2 \to \mathrm{B}_2$). 4 is obtained from 1, 2 and 3 by applications of Lemma 4 (Cut within belief). 5 follows from 4 by an application of Lemma 4. 6 is proved with the fact that $\neg\neg 2W$ is classically equivalent to $2W$ and an application of ($\mathrm{B}_2 \to \mathrm{B}_2$). 7 follows from 5 and 6 by Lemma 4.

Finally, 8 follows from 7, $\mathrm{B}_2\neg\mathrm{B}_1 1W$ in Γ_2',

$$\neg\mathrm{B}_1 1W, (\neg\mathrm{B}_1 1W \Rightarrow 2W) \to 2W,$$

an application of ($\mathrm{B}_2 \to \mathrm{B}_2$) and two cuts. This completes *Player 2* 's deduction.

As an example, we will present the part of the proof-figure which derives 8 from 7.

$$\dfrac{\overset{7}{\Gamma_2' \to \mathrm{B}_2(\neg\mathrm{B}_1 1W \Rightarrow 2W)} \qquad \dfrac{\dfrac{\neg\mathrm{B}_1 1W \to \neg\mathrm{B}_1 1W \quad 2W \to 2W}{\neg\mathrm{B}_1 1W, \neg\mathrm{B}_1 1W \Rightarrow 2W \to 2W}\,(\Rightarrow\to)}{\mathrm{B}_2\neg\mathrm{B}_1 1W, \mathrm{B}_2(\neg\mathrm{B}_1 1W \Rightarrow 2W) \to \mathrm{B}_2 2W}\,(\mathrm{B}_2 \to \mathrm{B}_2)}{\Gamma_2' \to \mathrm{B}_2 2W}\,(cut)$$

Note. In fact, we did not need all the formulas in Γ_2' for the conclusion of Proposition 3. Only the subset consisting of $\mathrm{B}_2\neg\mathrm{B}_1 1W, \mathrm{B}_2(\neg 2W \Rightarrow \mathrm{B}_1\neg 2W)$ sufficed.

5 Remarks

We have seen that even in a very simple puzzle, there are some ambiguities, and believe that such ambiguities are intrinsic in human logical activities. This may have some implications in economics and in other fields of science and engineering; see, for example, Oda and Yasugi [12]. To avoid to be involved too deeply in general arguments, however, we will conclude this article with some concrete remarks. For details, see [11], [12] and [15].

The notion of not knowing The notion of *not knowing* (not believing) plays a crucial role in the *two wise girls puzzle*. We interpreted the sentence "*Player 1* does not know (believe) A" as $\neg\mathrm{B}_1 A$. Nevertheless, there can be other possibilites of expressing *not know*. $\mathrm{B}_1\neg\mathrm{B}_1 A$ is a sensible alternative. In fact, we can replace $\neg\mathrm{B}_1$ by $\mathrm{B}_1\neg\mathrm{B}_1$ in the preceding arguments. Logically, the latter implies the former in $\mathrm{KD4}^2$.

Since $\neg\mathrm{B}_1$ suffices to our persent purpose, we do not go any further on this subject.

Alternative interpretation We can interpret our foregoing argument in a different manner.

Player 1 can claim with certainty that she does not know her hat is white if she is aware that $\Gamma_1 \to \mathrm{B}_1 1W$ cannot be proved within her logical system. *Player 1* can find out this unprovability if she can study the logical system in which she reasons from outside. She can definitely claim that she does not know she wears white if she can *jump out of* the system $\mathrm{KD4}^2$, borrowing an expression from Hofstadter [6].

Solvability of the puzzle We have assumed that, in answering the puzzle, both players are wise enough to reason logically. With regards to the interpretation of this article, such an ability suffices for the players.

With the alternative interpretation as mentioned above, we can go further and claim that the puzzle is *solvable*, that is, the following are all *determined automatically*.

Both Γ_1 and Γ_2 are consistent in $\mathrm{KD4}^2$.

Player 1 can show that there is no proof-figure of $\Gamma_1 \to \mathrm{B}_1 1W$ in $\mathrm{KD4}^2$.

Player 2 can show that Γ_2' is consistent in $\mathrm{KD4}^2$.

Player 2 can construct a proof-figure of $\Gamma_2' \to \mathrm{B}_2 2W$.

More precisely, there is an *algorithm* to evaluate the unprovability of each metatheoretic proposition (cf. Propositions 1 and 2), and there is also an *algorithm* to construct a proof-figure of *Player 2* 's conclusion (Proposition 3).

General case The essence of meta-theory of the wise girls puzzle has been fully explained with the case of the two girls puzzle. Although technically the proofs of the propositions in Sections 3 and 4 are much more complicated for the general case of *n* girls, there is no difference in the preceding observations.

References

1. Chellas, B.: Modal logic. Cambridge: Cambridge University Press 1980
2. Gentzen, G.: Investigations into logical deduction, The Collected Papers of Gerhard Gentzen, pp. 68–131. Amsterdam: North-Holland 1969
3. Gerbrandy, J.: Bisimulations on Planet Kripke. ILLC Dissertation Series, Universiteit van Amsterdam (1999)
4. Halpern, J.H., Moses, Y.: A guide to completeness and complexity for modal logics of knowledge and beliefs. Artificial Intelligence **54**, 319–379 (1992)
5. Hayashi, S.: Suri-Ronrigaku. Tokyo, Japan: Korona-sha 1989
6. Hofstadter, D.R.: Gödel, Escher, Bach. New York: Basic Books 1979
7. Hughes, G.E., Cresswell, M.J.: A companion to modal logic. London: Methuen 1974
8. Kaneko,M.: Introduction to epistemic logics and their game theoretic applications. Economic Theory (this issue, 2002)
9. Kaneko, M., Nagashima, T.: Game logic and its applications II. Studia Logica **58**, 273–303(1997)
10. Kleene, S.C.: Mathematical logic. New York: Wiley 1967
11. Oda, S.H., Yasugi, M.: Inference within knowledge. Discussion Paper Series 27, The Society of Economics and Business Administration, Kyoto Sangyo University; available at http://www.kyoto-su.ac.jp/~yasugi/Recent (1998)
12. Oda, S.H., Yasugi, M.: Jumping out from the system. The Proceedings of International Workshop on Emergent Synthesis, pp. 257–262; available at http://www.kyoto-su.ac.jp/~oda/projectsE (1999)
13. Ohnishi, M., Matsumoto, K.: Gentzen method in modal calculi. I. Osaka Math. J. **9**, 113–130 (1957)
14. Sato, M.: A study of Kripke-type models for some modal logics by Gentzen's sequential method. Publications of RIMS **13**, 381–468 (1977)
15. Yasugi, M., Oda, S.H.: A proof-theoretic approach to knowledge. Discussion Paper Series 29, The Society of Economics and Business Administration, Kyoto Sangyo University; available at http://www.kyoto-su.ac.jp/~yasugi/Recent (1999)