Jumping out of the system: an analysis by proof-theory
Pre-Forum seminars, May 22/2003, at KSU *

Mariko Yasugi Faculty of Science, KSU †
Sobei H. Oda Faculty of Economics, KSU ‡

1 Introduction: observing oneself from outside

In an intellectual activity, a human being thinks or infers not only within a fixed system but also occasionally takes thought of the system in which one dwells. Namely, a human being is endowed with the ability to observe oneself from outside. We interpret this action as **jumping out of a system**[1], and will make detailed analysis of such circumstances with a simple example. The basis of the analysis is a formal logical system as well as its proof-theory. [2] We will realize this purpose guided by an over-simplified puzzle. It does not mean, however, that we put importance on this particular puzzle. It is meant that with a simple example we can see heart of the matter without being dragged by its content.

The material of our discourse is a puzzle of guessing the color of a hat, known as "three wise men puzzle." [3] There are several versions of this puzzle. We define the puzzle in a way that is logically easier to manage.

Furthermore, since the number of players is irrelevant, we will deal with "two wise girls puzzle."

Two wise girls puzzle Two girls, Anna and Bianca, are seated in a row. An observer puts a hat on the head of each of them. It will be assumed that Bianca's hat is white, while Anna's hat can be of any color. Neither of the girls can see her own hat. Anna can see Bianca's hat, but not conversely.

*excerpt from "Journal of the Japan Association for Philosophy of Science" vol.28,no.1(2001), 87-92(in Japanese)
†yasugi@cc.kyoto-su.ac.jp. This work is partly supported by Scientific Grant-in-Aid (No.12878061).
‡oda@cc.kyoto-su.ac.jp. This work is partly supported by Scientific Grant-in-Aid (No.13480115).
[1] Hofstadter
[2] Proof-theory is a meta-study of a formal system.
[3] proposed by John McCarthy as a target of studies of artificial intelligence

The observer tells the girls: At least one of the hats on your heads is white.

The observer first asks Anna: Is your hat white?
(A) Anna answers: I don't know.
(B1) Bianca hears Anna's answer and adds it to her knowledge.
The observer next askes Bianca: Is your hat white?
(B) Bianca answers: Yes.

In order that one reach a right conclusion, one must be able to (logically) correctly reason. It must be, therefore, that the girls have inferring (reasoning) abilities in a common logical system if they can reach the correct conclusion through their communications. We can observe that, from the content of the puzzle, the process of Bianca reaching her conclusion (B) can be executed in a simple logical system with an operator which expresses "knowledge."

It should be appropriate to interpret the ground of Anna answering "I don't know" (that is, (A)) as follows: It is impossible for Anna to infer that she knows she wears white whithin the system in which she dwells. We should interpret the fact that Anna can recognize this impossibility as a consequence of her ability to jump out of her own inference system and observe "about the system" from outside. We are thus necessarily led to the notion of "jumping out of a system."

We call the action of jumping out of a sytem and observing the system "meta-inference" or meta-reasoning. Bianca in (B1) also needs to jump out of her system. We should therefore assume that two girls have an equivalent ability of meta-resoning.

As a matter of fact, the action of taking the result of meta-reasoning back to one's own system is also necessary. It is hence desired that jumping out of and returning to a system can be treated with a unified method.

Usually, one would assert that "a puzzle has been solved" with (B). If the ability to jump out of a system is of transcendental nature, however, Anna would not be able to claim with confidence that she "does not know." Similarly, if the proof construction of the affirmative solution (B) is a gift of Bianca's capability, we cannot claim the "solvability of the puzzle" in the sense that "it can be solved by anybody." In order that a correct conclusion can be attained without depending on individual abilities of observing one's own system from outside and proving a proposition in a system, it will be appropriate to endow the process of reaching the solution with decidability or automatism.

In order to deal with these circumstances rigorously, we have defined an appropriate logical system, and formulated meta-observation about the system in terms of "proof-theory." We will attempt to throw light on the necessity and appropriateness of our attempt through detailed analysis.

We would like to note that there are hundreds of articles on wise men puzzle, or muddy children puzzle for that matter. As far as we know, they assume formal systems from the start and treat them with semantics.

2 Analysis of the process of puzzle-solving

We will use the following notations. Γ and Δ denote respectively finite sets of propositions (knowledge), say $\Gamma : A_1, A_2, \cdots, A_m$, and $\Delta : B_1, B_2, \cdots, B_n$. $\Gamma \rightarrow \Delta$ will denote that the conjunctive premise Γ, that is, $A_1 \wedge A_2 \wedge \cdots \wedge A_m$ implies the disjunctive conclusion Δ, that is, $B_1 \vee B_2 \vee \cdots \vee B_n$.

We will analyze two girls' mental activities.

A-1. Anna's initial knowledge set $\mathcal{A}_1$: "Anna knows that at least one of the hats is white ($K_1(1W \vee 2W)$)", "Anna knows that Bianca wears white ($K_1 2W$)"

Anna tries to prove

A-2. $\mathcal{A}_1 \rightarrow$"Anna knows that she wears white ($K_1 1W$)."

She then realizes that

A-3. $\mathcal{A}_1 \rightarrow$"Anna knows that she wears white ($K_1 1W$)" cannot be inferred (derived) within Anna's inference system.

This entire cognition cannot be inferred within the system, but can be obtained by an observation of the system from outside.

We call a statement like A-3 a "meta-theorem" (a theorem about the system), when it holds. Anna next translates the meta-theorem A-3 as "Anna does not know that she wears white ($\neg K_1 1W$)", returns to her own system with the proposition (statement) $\neg K_1 1W$, and then expands her knowledge set by adding the new statement to it. Let us call the process of translating a meta-theorem to a statement in the language of a system a "translation-rule." So, Anna has applied a translation-rule to her meta-theorem. She then tests if her expanded knowledge set is consistent, that is,

A-4. $\mathcal{A}_2 : \mathcal{A}_1, \neg K_1 1W \rightarrow \perp$ cannot be inferred in the system, where $\perp$ exresses a contradiction. This is another meta-theorem.

With the expanded knowledge set, Anna can trivially infer $\neg K_1 1W$:

$$\mathcal{A}_1, \neg K_1 1W \rightarrow \neg K_1 1W$$

and she claims "I don't know that I wear white."

Bianca's initial knowledge set can be expressed as $\mathcal{B}_1 : K_2(1W \vee 2W)$. Upon hearing Anna claiming $\neg K_1 1W$, she understands it as $K_2 \neg K_1 1W$. This is another translation-rule. She then adds to $\mathcal{B}_1$ the new statement $K_2 \neg K_1 1W$ to obtain her new knowledge set $\mathcal{B}_2 : \mathcal{B}_1, K_2 \neg K_1 1W$. She next tests if $\mathcal{B}_2$ is consistent, that is,

B-1. $\mathcal{B}_2 \rightarrow \perp$ is not inferred in her system.

Securing the consistency of $\mathcal{B}_2$, Bianca finds that she wears white, that is she infers in her system

B-2. $\mathcal{B}_2 \rightarrow K_2 2W$.

Summing up, the solvability of the two wise girls puzzle involves the following.

(1) Expression of statements (propositions) (2) Players' knowledge sets (3) Reasoning systems (4) Proof (derivation) within a system (5) Jumping out of and returning to a system (6) Translation-rule (7) Expansion of a knowledge set (8) Automatism of non-derivability (Decision algorithm of A-3, A-4, B-1) and derivability in a system (Construction algorithm of a derivation of B-2 within the system)

3 Formal treatment

We will subsequently give a brief account of each of (1)∼(8).

(1) The formulas of propositional logic with two knowledge operators K_1 and K_2.

(2) A knowledge set Γ is a consistent and finite set of formulas either enclosed with the knowledge operators or their negations. (A finite set of formulas Γ is called consistent if $\Gamma \to \perp$ is not provable in KD^4.) So, in fact Anna and Bianca should have started by checking if their initial knowledge sets were consistent.

(3) Propositional logic with the following axioms for knowledge operators added:

$$K_i X \to K_i K_i X \ (i = 1, 2)$$

This axiom is known as **4** and the system is known as the system of modal logic KD^4. The axiom **4** is also known as one of the axioms of bounded rationality defined by Rubinstein. (This corresponds to the information structure where $\omega' \in P(\omega)$ implies $P(\omega' \subset P(\omega))$.)

(4) Inferences within KD^4. When we say a statement of the form $\Gamma \to \Delta$ can be inferred (derived), we mean that there is a proof (derivation) of it within the system KD^4. We say $\Gamma \to \Delta$ is provable (derivable) in KD^4.

(5) Demonstration of formal unprovability (non-derivability). This is a meta-theorem.

(6) This part is our interpretaion. When $\mathcal{A}_1 \to K_1 1W$ is shown to be unprovable, we translate it to $\neg K_1 1W$ (under assumption $\mathcal{A}_1$). When Bianca recognizes Anna's claim $\neg K_1 1W$, we translate it to $K_2 \neg K_1 1W$. Here we need some kind of transfer principle.

(7) In expanding a knowledge set by adding a new formula, one has to make sure the expanded set is consistent. This is a meta-theorem.

(8) There is an algorithm to decide whether a statement $\Gamma \to \Delta$ is provable (in which case a formal proof can be automatically constructed) or is unprovable. This can be done by proof-theory on the system. Namely, any formal proof can be transformed into a "normal form" of the same conclusion which has no detour. So, if a statement can be proven in the system, it can be proven with a normal proof. With a normal proof, there is an algorithm as required above.

There is an algorithm to construct a possible normal proof of a given statement. The process eventually terminates, producing a figure. The statement is provalbe if and only if the resulting figure is a correct normal proof.

We can thus conclude that the two wise girls puzzle is solvable.

The mathematical treatment of (8) is seen in our paper in Economic Theory (cf. Proceedings); more speculation on the knowledge operator and related axioms is seen in our paper of SCMJ (cf. Proceedings).