S5 solution to the red hat puzzle

Robert C. Robinson[1]
Florida State University

Abstract: I argue that the solution to the Red Hat Problem, a puzzle derived from interactive epistemic logic, requires *S5*. Interactive epistemic logic is set out in formal terms, and an attempt to solve the red hat puzzle is made in $K_{\tau\sigma}$, $K_{\rho\tau}$, and $K_{\rho\sigma}$, each of which fails, showing that a stronger system, $K_{\tau\sigma\rho}$ is required.

1. Interactive knowledge

Interactive epistemology is just what it sounds like: knowledge about others' knowledge, as in 'I know that you know that I know…'. Epistemic logic goes back to Hintikka 1962, but the formal analysis of interactive epistemology starts with Aumann 1976. The formal system I describe follows Aumann, especially in Aumann 1999a and Aumann 1999b.

So formally, let Ω be the set of all *states of nature*, or states $w = (w_1, w_2, \ldots, w_n)$. Ω consists of many partitions, each of which is said to be the information set of an individual $i = (1, 2, \ldots N)$. An event E is a subset of Ω, and is said to hold (or obtain) at w iff $w \in E$.

We can introduce a knowledge operator, K, which operates on Ω, into the system. Call K_i the *individual knowledge operator* of agent i. Let K_iE be the event described by 'agent i knows E'. Note that K_iE is itself an event, and so is a subset of Ω. Therefore, K_iE obtains at some w, and does not obtain at others.

We can also introduce a *knowledge function* k on Ω. k is a set of different 'states of knowledge' for some individual i. $k(w)$ represents the knowledge i possesses when w is the true state of the world.

Agent i is said to *know* E at w iff E includes his information set at w. Call $I(w)$ agent i's information set. Formally,

[1] I wish to acknowledge the comments of Piers Rawling on an earlier draft of this paper, as well as the comments of three anonymous referees, particularly in regards to my notation.

$I(w) = \{w' \in \Omega: k(w) = k(w')\}$

which says that *i*'s information set includes all the worlds *w* that *i* cannot differentiate from one another. $I(w)$ consists of all the states *w*' that *i* considers to be possible. For any *w* and *w*', $I(w)$ and $I(w')$ are either disjoint or identical.

Finally, knowledge operators can operate on one another (since applications of a knowledge operator to an event are events), so K_iK_jE says that '*i* knows that *j* knows that *E* obtains'.

See van der Hoek and Verbrugge 2002 for a discussion of the axiomatization of the system, especially § 2.4. See Aumann 1999a for a discussion of the relationship between the semantics and syntax of the system.

2. Common knowledge

Interactive knowledge provides a challenge for philosophers in at least one immediate way — i.e., *common knowledge*.

Let us begin with the following scenario. Three schoolgirls are sitting in a circle, so that each can see the other two. Each is wearing a hat, and each knows that her hat is either white or red. Each girl can see the colour of everyone else's hat, but not her own. And in fact, everyone's hat is red.

Each girl must guess her hat colour. If she guesses correctly, she may go to recess (Π=r). If she guesses incorrectly, she is assigned a particularly tedious homework assignment (Π=-h). If she does not guess at all, she must continue to sit in the room (Π=0).[2]

The game proceeds sequentially[3]: The teacher asks each girl if she can identify the colour of her own hat. Of course, each says no in turn. Now the teacher happens casually to mention that at least one girl in the circle is wearing a red hat, a fact that does not teach any girl any obviously new fact, since each can see that there are at least two red hats in the circle. Nonetheless, when the teacher again asks the first girl if she knows her hat colour, she replies 'no,' and when

[2] Where h > r > 0 > -h.

[3] Note that van Benthem 2004 describes a similar game, with the exception that his is a simultaneous move game, and so common knowledge and common knowledge of rationality are derived differently.

she asks the second girl if she knows her hat colour, the reply again is 'no.' But when the teacher asks the third girl, she correctly answers 'yes, I am wearing a red hat.'

The puzzle really is that, besides mentioning some innocuous detail that each girl already knew, no bit of information changed from the first round of questions to the second. And so how did the third girl manage to deduce her hat colour based on this new information, when actually the new information did not teach her anything she did not already know?

The answer, of course, is that the girl *did* learn something new — not that there was at least one red hat, but that *every girl knows that every girl knows* that there is at least one red hat. We call this common knowledge.

Our definition of common knowledge comes from Lewis 1969: E is common knowledge iff everyone knows E, everyone knows that everyone knows E, and so on, *ad infinitum*. More strictly, we can define a *common knowledge operator* such that: E is common knowledge for agents ($i \ldots n$) in I, iff for all i, k_iE, k_jE, k_ik_jE, and so on. Let K^1E represent the event that everyone knows, and call E mutually known. Let K^mE mean that E is known to the m^{th} degree, where k_ik_jE is K_2E. E is common knowledge iff $K^\infty E$.[4]

We are now in a better position to analyze the three hats problem.[5] On the first round, each girl's (call them, colourfully, One, Two, and Three, respectively) partition of Ω represents her knowledge of the state of nature w. For example:

$$I_3(w) = \{RRR, RRW\}$$

which says that Three's information set includes the world in which One and Two have a red hat and Three has a red hat, and the world in which One and Two have a red hat and Three has a white hat. She cannot differentiate between those worlds. One and Two have

[4] We may make even more fine grained distinctions between e.g., mutual, distributed, individual, common, and 'almost' common knowledge — (cf. Koessler 2000).

[5] The three hats problem is derived from the 'Dirty Faces Problem', first described in Littlewood 1953, in which each of three women on a train can see each other woman's face, but not her own, each woman's face is dirty, and each woman laughs at the others until realizing her own face is dirty.

similar information sets, representing their own ignorance of their own hat colour.

However, the teacher's announcement of a fact already known to Three adds some bit of knowledge — i.e., common knowledge of at least one red hat. She is now in a position to deduce her own hat colour using her knowledge of the accessibility relations of transitivity, symmetry, and reflexivity.[6]

Weber 2001 makes explicit the importance of common knowledge of other factors, i.e., rationality of all players. In most other strategic games (such as the Ultimatum Game), especially those used to study iterated reasoning (such as a p-Beauty Contest game), players would prefer to be 'smarter' than other players (cf. Nagel 1995 on the guessing game). This is not true here, however, since the strategy employed by Three depends on common knowledge of rationality of each player.

Now when the teacher asks again, One admits that she does not know her own hat colour. Now Two knows that *one hat is red* is common knowledge, and she knows her own information set, but is unable to derive her own hat colour. Three now has even further knowledge gained in the last round. She knows *one hat is red* is common knowledge, she knows One and Two's information partition, and she knows her own information partition. With this knowledge, she is able to derive her own hat colour.

If Two and Three's hats were white, then One could deduce her own hat colour. But she could not. This information reveals to Two and Three that at least one of them is wearing a red hat. Two then admits that she does not know her own hat colour. Three then notes that if her own hat had been white, then Two would have known that her own hat colour is red, since they both know that there's at least one red hat. But she could not. So Three correctly reasons 'my hat must therefore be red.'

See van der Hoek and Verbrugge 2002 for an engaging formal proof, and a similar intuitive and formal proof in Geanakoplos 1992, as well as in my own Robinson 2006.

[6] The interested and technically minded reader should refer to Kripke 1959 for details.

3. *S5*

I will argue that the solution depends on treating the relationship between the worlds in each girl's information partition as being symmetric, transitive, and reflexive.

Initially, the relevant parts of the information partitions of the girls looks as in Figure 1. This says that One knows her information partition, knows that Two's information partition contains at least 2 of the 4 worlds (but does not initially know which two), and knows that Three's information partition contains at least 2 of the 4 worlds (but again, does not know which two). The same reasoning applies to the other two girls.

One	↗	$I_1(w)$
	→	$I_1(w) \subseteq K_2$ {*RRR, WWR, RWR, WRR*}
	↘	$I_1(w) \subseteq K_3$ {*RRR, WRR, RRW, WRR*}
Two	↗	$I_2(w) \subseteq K_1$ {*RRR, RWR, WRR, WWR*}
	→	$I_2(w)$
	↘	$I_2(w) \subseteq K_3$ {*RRR, RRW, RWR, RRW*}
Three	↗	$I_3(w) \subseteq K_1$ {*RRR, WRR, RRW, WRW*}
	→	$I_3(w) \subseteq K_2$ {*RRR, RWW, RWR, RRW*}
	↘	$I_3(w)$

Figure 1: Initial Information Partition

When One admits that she does not know her hat colour, Two's information partition changes to include $I_2(w) \subseteq K_1$ {*RRR*, *RWR*, *WRR*} (i.e., it no longer contains {*WWR*}). This is not enough to allow her to determine her own type.

Beginning at Three's turn, we see the knowledge tableau, which is represented in Figure 2.

Fact. The solution to the Three Hats Puzzle requires nothing less than an S5 (also known as $K_{\rho\tau\sigma}$) interactive modal epistemic logic. Reflexivity, symmetry, and transitivity are two part relations. In this case,

treat *information partitions* as the relata, and *'has epistemic access to'* as the relationship.

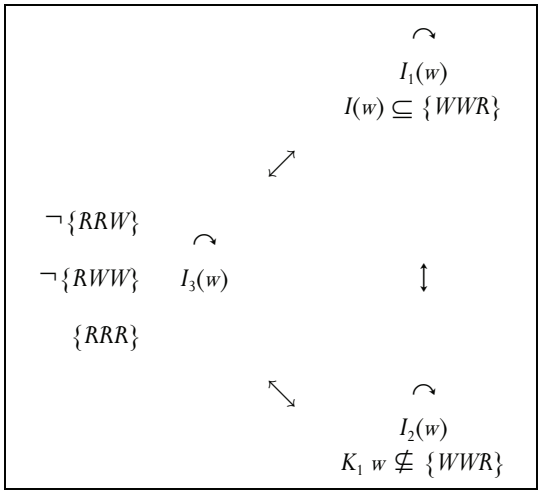

Figure 2: Final Knowledge Tableau

Proof. Assume $K_{\tau\sigma}$. Three would be able to learn that $K_2K_1 \nsubseteq \{WWR\}$, but would not be able to look to $I_2(w)$. Three determines her own type by eliminating $\{WRR\}$ from her information partition.

Assume $K_{\rho\tau}$. Without symmetry, the strong requirement of common knowledge of rationality is not satisfied.

Assume $K_{\rho\sigma}$. When $n \leq 3$ players, $K_{\rho\sigma}$ is equivalent to $K_{\rho\tau\sigma}$. Since if each girl has access to both other girls, then each girl has access to every girl each girl has access to. When $n \geq 4$ (e.g., if there were at least 4 girls) Four would need to know what Three knows about Two, etc. Four's access to Three is not sufficient. Four must also know what Three knows about Two. This argument generalizes to all players. □

Robert C. Robinson
Dept. of Philosophy
Florida State University
151 Dodd Hall
Tallahassee FL 32306-1500
bcr2925@fsu.edu

References

Aumann, R. 1976. Agreeing to disagree. *Annals of Statistics* 4:1236–1239.

Aumann, R. 1999a. Interactive Epistemology I: Knowledge. *International Journal of Game Theory* 28:263–300.

Aumann, R. 1999b. Interactive Epistemology II: Probability. *International Journal of Game Theory* 28:301–14.

Geanakoplos, J. 1992. Common Knowledge. *Journal of Economic Perspectives* 6(4):53–82.

Hintikka, J. 1962. *Knowledge and Belief*. New York: Cornell University Press.

Koessler, F. 2000. Common Knowledge and Interactive Behaviors: A Survey. *European Journal of Economic and Social Systems* 14(3):271–308.

Kripke, S. 1959. A Completeness Theorem for Modal Logic. *Journal of Symbolic Logic* 24:1–14.

Lewis, D. K. 1969. *Convention: A Philosophical Study*. Cambridge, MA: Harvard University Press.

Littlewood, J. 1953. *A Mathematician's Miscellany*. London: Methuen and Company Limited.

Nagel, R. 1995. Unraveling in Guessing Games: An Experimental Study. *The American Economic Review* 85(5):1313–1326.

Robinson, R. C. 2006. Bounded Epistemology. *SSRN eLibrary*. http://ssrn.com/paper=1000697.

van Benthem, J. 2004. What One May Come To Know. *Analysis* 64:95–105.

van der Hoek, W. and Verbrugge, B. 2002. Epistemic Logic: A Survey. In Petrosjan, L. and Mazalov, V., eds., *Game Theory and Applications,* vol. 8, pp. 53–94. New York: Nova Science Publishers.

Weber, R. 2001. Behavior and Learning in the 'Dirty Faces' Game. *Experimental Economics* 4:229–42.