Common Knowledge and the Game of Red Hats

PETER NEWMAN

The Johns Hopkins University, Baltimore

1. THE GAME

Once upon a time a cruel king decided to have some sport with the men incarcerated in his dungeons. Commanding his guards to blindfold them and to bring them into the Great Hall of the castle, he addressed the prisoners as follows:

'We are going to play a little game. Each of you will have a hat placed upon his head. Some hats may be blue and the rest will be red; otherwise, the hats are identical. Your eyes will then be uncovered and you will be free to look at your fellow prisoners and their hats. However, on penalty of instant beheading, at no time will any of you communicate in any way with his fellows, nor will he look up to ascertain the colour of his own hat.

A gong will then sound, at which time each of you is to raise his hand if he believes that his own hat is red. A correct guess will earn him instant freedom. An incorrect guess, however, will bring him instant execution.

After that a bell will ring, whereupon those of you still left will take a short break. Then the gong will sound again and a second round will be played, with the same rules as before. The game will continue thus until I become bored or none of you is left, whichever comes first.'

Let us say that this cruel but interesting game has a solution if it is possible for the prisoners to act in such a way that all those with red hats go free and none is executed. Whether the game has such a solution turns out to depend on how many red hats there are, on the type and amount of the information available to the prisoners, and—more surprisingly—on the degree to which the prisoners hold that knowledge in common. The discussion which follows analyzes how the possibility of a solution changes as each of these three factors is varied.

This game of red hats is said to be 'well-known', i.e. its origins are obscure. I first heard of it from Edi Karni, and read of it in a manuscript by Sérgio Ribeiro da Costa Werlnag. I am much indebted for further enlightenment to discussions with Steve Blough, Loie Harrington, Hiro Kawai and Hugh Rose.

Every fable has a moral. This fable is peculiarly well designed to bring out the moral that how a group knows a fact may be just as important as what that fact may be. The theorem suggested in Section 4 (which is probably also 'well-known') says that in this particular version of the fable the moral takes the form of a quite precise trade-off between the what and the how.
2. INFORMATION; AND THE CASE OF NO RED HATS

(a) A priori the prisoners, say n in number, know only that the number m of red hats belongs to the finite set of integers \( \{0, 1, 2, \ldots, n\} \). Assume that the prisoners’ guards, out of pity or terror or even bribery, decide secretly to whisper truthful new information to the prisoners. This information could of course be complete: ‘You, prisoner Smith, are wearing a red hat.’ It could be irrelevant: ‘The Queen Mother is wearing a red hat.’ Suppose that the guards are neither brave enough to give complete information nor sadistic enough to impart irrelevant facts. It then seems natural to restrict the type and amount of information available here to the situation where the guards are willing to reveal only news which effectively reduces the possibilities, i.e., to statements of the form ‘\( m \geq j \)’, for \( j = 1, 2, 3, \ldots, n \).

Notice that if the information is to be new then \( j \) must be positive. That \( m \geq 0 \) can be worked out for himself by any prisoner. Moreover, everyone can figure out that everyone can figure out . . . that \( m \equiv 0 \). But, no matter how generally known, this no news is good news only in the sad case when there never are any red hats anyway, when \( m = 0 \) and the king cheats.

(b) However, it will do the tyrant no good to cheat in this way. At the first round each prisoner is in a state of complete nescience, seeing no red hat nor knowing anything other than \( m \equiv 0 \). Aware of the mortal penalty for wrongly raising his hand, each prisoner will therefore pass. This general inaction will yield no new information for the second round, and so everyone will pass again; and again, and again, . . . A good thing too, since after all no one has a red hat and so under the King’s rules none can win his freedom. The most the prisoners can hope for is that ennui will overtake the monarch before they themselves die of exhaustion. This sad outcome still fits the definition of a solution, however, so that oddly enough the degenerate case \( m = 0 \) can always be solved.

3. THE THREE CASES OF RED HATS

(i) The case \( m = 1 \), in which the one red hat is worn by the prisoner Smith

(a) Suppose first that, as before, each prisoner knows only that \( m \geq 0 \). Everyone but Smith will see his red hat and Smith himself will see none. But none of the prisoners, not even Smith, can be sure that his own hat is red; so, as before, each will pass at every round. Since Smith actually has a red hat and could go free if he guessed right, this means that the game now has no solution.

Provided the information available to each prisoner remains at \( m \equiv 0 \), the game continues to be insoluble no matter how large \( m \) may be. Even if each prisoner has a red hat he will see only \( (n - 1) \) of them, and so cannot be sure that his own hat is red. Everyone will pass at every round and none of the red-hatted prisoners will gain his freedom.

(b) Suppose next that before the first round the guards whisper to each man separately that there is at least one red hat, i.e., that \( m \geq 1 \). This new information is private to each prisoner, not held in common between them. Nevertheless it suffices to solve the game, for Smith can argue as follows:

‘I know that there is at least one red hat but I see none. So it must be on me. Ergo, I will raise my hand at the first round.’

Smith’s argument is correct and he will go free. Every other prisoner will see Smith’s red hat, but knowing only that \( m \geq 1 \) may suspect his own head to be red-less, and so will pass at the first round. Further, so they all pass the game will ha

(ii) The case \( m > 1 \)

(a) If the above pass at all rounds Smith will see j red hats and could go free if he guessed right, this means that the game now has no solution.

(b) Next, suppos that each prisoner the of rational agents

‘I know that the j red hats are not mine,’ says Smith, being at so that at round or round, and the ga

Observe that the latter will occur again a

(c) In this next ‘\( m \equiv 1 \)’ to each pr being told the same

Informed that even

Such informati definition become been realized first of justice and pro Convention (1969).

The first round second round Smi

‘I see only o possibly—m

see precisely knows that \( i \) have reasone

‘I, Jones, I will rais So he mu
m of red hats belongs s, out of ploy or terror t the prisoners. This \textit{red hat}. It could be wards are neither brave nt facts. It then seems te situation where the ites, i.e. to statements

\[m \geq 0\] can be worked t everyone can figure good news only in the ng cheats.

round each prisoner ther than \(m \geq 0\). Aware fore pass. This general ne will pass again; and nd so under the King's ut \textit{ennui} will overtake ill fits the definition of always be solved.

\begin{quote}
Everyone but Smith will not even Smith, can be since Smith actually has now has no solution.
\end{quote}

\textit{The Game continues to be:} hat he himself will see yone will pass at every nan separately that there ach prisoner, not held in ith can argue as follows: t be on me. Ergo, I will

will see Smith's red hat, d so will pass at the first

round. Further, Smith's success at round one will not improve their information at round two, and so they all pass then too; and indeed at every subsequent round. Thus, according to the definition, the game will have a solution.

(ii) \textit{The case \(m = 2\), in which the two red hats are worn by the prisoners Smith and Jones}

(a) If the available information remains as in (b), i.e. just the statement \(m \geq 1\), everyone will pass at all rounds. Before the first round everybody except Smith and Jones will see two red hats. Smith will see Jones's, and Jones will see Smith's. Since everyone knows only that \(m = 1\), none can be quite sure that what sits on his own head is red. So they will all pass at the first round. The second and subsequent rounds bring no further information, and everyone will continue to pass. Smith and Jones will not go free, everyone will remain prisoner, and the game will have no solution.

(b) Next, suppose that the guards reduce the possibilities to \(\{2, 3, 4, \ldots, n\}\) by whispering to each prisoner the information \(m \geq 2\). Thus, before the first round Smith can (and so, in this model of rational agents, will) argue as follows:

'I know that there are at least two red hats but I see only one, on Jones. So the other must be on me. Ergo, I will raise my hand at the first round.'

Jones, being as bright as Smith and having the same information, will reason in the same way, so that at round one both will raise their hands and go free, the other prisoners will all pass at every round, and the game will be solved. Essentially, the game will be over at the first round.

Observe that Smith's reasoning here is the same, \textit{mutatis mutandis}, as his reasoning in (b). Call the latter reasoning of Type A1, and that here of Type A2. Reasoning of this generic Type A will occur again and again.

(c) In this next sub-case assume that the guards revert to whispering their earlier statement \(m \geq 1\) to each prisoner, but that in addition they also tell him that all his fellows are individually being told the same thing; however, they do not tell him that any prisoner besides himself is being informed that everyone knows that \(m = 1\).

Such information as there is now on the size of \(m\) stops being private and by (a temporary) definition becomes common knowledge, an important concept whose significance seems to have been realized first by David Hume (in \textit{A Treatise of Human Nature}, 1739, III.i.2, 'Of the origin of justice and property'), and to have been analyzed first by David Lewis in his luminous book \textit{Convention} (1969).

The first round now proceeds as in 2 (a); everybody passes. But during his break before the second round Smith, if he is reasonably bright, will reason as follows:

'I see only one red hat, perched atop Jones. So there can be at most two red hats, Jones's and possibly—mine. So Jones himself cannot see two red hats. Moreover, I can infer that he can see precisely one. For suppose not, i.e. suppose that he sees no red hats. I know that Jones knows that \(m = 1\), so I know that before the first round he, being as bright as I am, would have reasoned as follows:

"I, Jones, know that there is at least one red hat but I see none. So it must be on me. Ergo, I will raise my hand at the first round". Now \textit{continues Smith}, Jones did no such thing. So he must see just one red hat.'

\textit{q.e.d.}
But I, Smith, also see just one red hat, and that on Jones himself. Hence, the only person that Jones can see with a red hat must be me. Ergo, I will raise my hand at the second round.

Jones is as bright as Smith and has similar information, though Smith does not know this. Hence, he will reason in the same way, so that at the second round both will raise their hands and go free. The other prisoners will all pass at every round, and the game will be solved.

It is the commonality of the information that \( m \geq 1 \) which forces a solution in this subcase, even though for \( m = 2 \) the statement \( 'm \geq 1' \) is strictly less informative than \( 'm \geq 2'. \) Moreover, the assumption of common knowledge enables each red-hatted prisoner to presume before round two that the other would have already employed Type A reasoning before round one.

(iii) The case \( m = 3 \), in which the red hats are worn by the prisoners Smith, Jones and Brown

(a) Go back now to the situation in which information is purely private and suppose that it consists of either \( 'm \geq 1' \) or \( 'm \geq 2'. \) Then an argument similar to that of 2 (a) will show that everyone will pass at every round. The game has no solution.

(b) Suppose next that information is still given privately to each prisoner but consists now of the statement \( 'm \geq 3', \) thus reducing the possibilities to \( \{3, 4, 5, \ldots, n\} \). Then before the first round Smith can argue as follows:

'I know that there are at least three red hats but I see only two, on Jones and Brown. So the other must be on me. Ergo, I will raise my hand at the first round.'

Jones and Brown, being as bright as Smith and having the same information, will reason in the same way so that at the first round all three will raise their hands and go free, the other prisoners will all pass at every round, and the game will be solved. Notice that Smith's reasoning is again of type A, and that again the number of red hats in the first sentence is increased by one, compared with 2 (b); so call this reasoning to Type A2.

(c) Suppose next that the information whispered by the guards to each prisoner is the compound statement \( 'm \geq 2' \) and we are telling each of your fellow prisoners the same thing. Reasoning as in 2 (b) it is easy to conclude that in this situation everyone will pass at the first round. However, during his break before the second round Smith will argue as follows:

'I see only two red hats, perched atop Jones and Brown. So there can be at most three red hats, Jones's, Brown's and—possibly—mine. Hence, Jones himself cannot see three red hats. Moreover, I can infer that he can see precisely two. For suppose not, i.e. that he sees only Brown's red hat. I know that Jones knows that \( m \geq 2 \), so I know that before the first round he, being as bright as I am, would have reasoned as follows:

"I, Jones, know that there are at least two red hats but I see only one, on Brown. So the other must be on me. Ergo, I will raise my hand at the first round."

Now [continues Smith] Jones did no such thing. So he must see two red hats.

q.e.d.

But I, Smith, also see two red hats, those on Jones and Brown. So the only two persons that Jones can see with red hats must be Brown and me.

I can give exactly the same rationale for Brown's failure to raise his hand at the first round;
PETER NEWMAN

the only person that
he second round.

he, hence, go free.

this subcase, even
2. Moreover, the
1 before round two.

Jones and Brown
and suppose that it
1 will show that
but consists now of
before the first round
s and Brown. So the
m, will reason in the
the other prisoners
's reasoning is again
ed by one, compared
ner is the compound
thing'. Reasoning as
first round. However,
be at most three red
not see three red hats.
i.e. that he sees only
before the first round
he, on Brown. So the
1 hats. q.e.d.
only two persons that
and at the first round;

Common Knowledge and the Game of Red Hats

so he also must see just two red hats, on Jones and me. Ergo, doubly assured, I will raise my
hand at the second round.'

Jones and Brown are as bright as Smith and have the same information, though he does not
know that. So each will reason in the same way, and at round two all three will raise their hands
and go free. The other prisoners will all pass at every round, and the game will be solved.

Once again the smaller amount of knowledge which each prisoner has is compensated for by
the greater degree of commonality in which everyone holds it. Smith's reasoning in this situation
is the same, mutatis mutandis, as his reasoning in 2 (c). Calling it of generic Type B, denote by
Type B2 Smith's reasoning in 2 (c) and by Type B3 his reasoning here. An integral part of Smith's
argumentation here is the assumption of common knowledge, by which he infers that both Jones and
Brown use Type A2 reasoning before round one, whereas in 2 (c) Smith infers that at that stage
Jones uses Type A1 reasoning. A useful shorthand notation for these two situations is that in 2 (c)
Smith's reasoning takes the form B2 → A1, whereas in 3 (c) it is of the form B3 → A2.

(d) Quite surprisingly, the interplay between the amount of each prisoner's knowledge and the
level of commonality at which it is held can be carried further. Suppose now that the quantitative
part of the information which the guards give to each prisoner is only the statement 'm ≥ 1' (rather
than 'm ≥ 2' or 'm ≥ 2'), but that in addition they tell him both that every prisoner knows that
m ≥ 1, and that every other prisoner is also being informed that every prisoner knows that m ≥ 1.

It is easy to check that now there is so little quantitative information available that everyone,
including Smith, Jones and Brown, will pass not only at the first round but also at the second. For
example, before each of the first two rounds Smith will see only two red hats and so he cannot
dismiss the possibility that there are no more; so he passes each time.

But during his break before the third round Smith can begin to reason to some effect, provided
he possesses the ratiocinative powers of Sherlock Holmes at work on a three-pipe problem. Smith
will reason as follows.

'I see only two red hats, perched atop Jones and Brown. So there can be at most three red
hats, Jones's, Brown's, and—possibly—mine. Hence, Jones himself cannot see three red
hats. Moreover, I can infer that he can see precisely two. For suppose not, i.e. that he sees
only Brown's red hat. I know that he knows that Brown knows that m ≥ 1, so I know that
before the second round Jones will see at most three red hats, and he would have reasoned as follows:

"I, Jones, know that there is at least one red hat and I see only one, on Brown. However,
there must be another. For suppose not. Then since I know that Brown knows that m ≥ 1,
I know that before the first round Brown, being as bright as I am, would have argued as
follows:

""I, Brown, know that there is at least one red hat but I see none. So it must be on me.
Ergo, I will raise my hand at the first round."

Now (reasons Jones [according to Smith]), Brown did not such thing. So he must see a
red hat. But I, Jones, also see a red hat, on Brown himself. So the only person that Brown
can see with a red hat must be me. Ergo, I will raise my hand at the second round."

But [we are back now to Smith's direct reasoning], Jones did no such thing. So he must see
exactly two red hats... q.e.d.
It is the case [continues Smith] that I also see two red hats, on Jones and Brown. So the only two persons that Jones can see with red hats must be Brown and me, Smith. I can also give exactly the same rationale for Brown's failure to raise his hand at the second round, so he must see Jones and me with red hats. Ergo, doubly assured, I will raise my hand at the third round.'

Jones and Brown are as bright as Smith and have the same information, though he does not know that. So they will reason in the same way, and at the third round all three will raise their hands and go free. The other prisoners will all pass at every round, and the game will be solved.

This new deeper level of common knowledge, what might be called common knowledge at the second level; again compensates for the further reduction in the amount of the available quantitative information.

Notice that the reasoning of all three red-hatted prisoners before round three is of a new type, say Type C3, which on the basis of assuming common knowledge at the second level presumes that, before the second round, all three prisoners employed arguments of Type B2, and this in turn presumed on the basis of (simple) common knowledge that before the first round everyone used arguments of Type A1. This state of affairs can be symbolized in the shorthand notation introduced in the discussion of 3 (c) by C3 → B2 → A1.

4. A THEOREM

The discussion so far has revealed a pattern in the interplay between the quantity of the knowledge available and the level at which it is held in common, a pattern whose nature will become apparent with the introduction of some further notation. What so far has been called simply common knowledge will henceforth be known as commonality at the first level, or CK1; there has just been introduced commonality at the second level, CK2; and so on. Thus, the situation where everyone knows that everyone knows that everyone knows . . . , that phrase is repeated i times in all, will be called commonality at level i, written CKi.

Purely private information, where common knowledge is quite absent, was introduced in 1 (b). It is useful to refer to this as a case of commonality at the zeroth level, written CK0. It is also helpful to remark that the assumed structure of this particular game ensures that if for some positive integer i quantitative information is held at level CKi by any prisoner, then for any non-negative integer k < i the same information is held at level CKk by every prisoner. Lewis (op. cit., pp. 59-60) shows that such a relation cannot be presumed for arbitrary games or situations.

A review of the examples given so far suggests the plausibility of the Red Hat Theorem

The Red Hat Theorem

Let there be n prisoners and m red hats, where m = 0, 1, 2, . . . , n. Let the quantitative information available to each prisoner be the statement 'm ≥ j', where j = 1, 2, . . . , m; and let this information be of commonality CKi, where i = 0, 1, 2, . . . Then:

(a) For m = 0 and any CKi, the game has a solution.

(b) For m > 0 the game has solution if i ≥ m - j.

(c) For m ≥ 0 if the game has a solution it will be found around number m - j + 1.

A proof of (a) was given in 2 (b). A heuristic 'proof' of (b) is as follows: From the remark above it suffices to prove that a solution exists if CK1 is actually such that i = m - j. Using the shorthand notation introduced to i = 1, etc. Ap contained in Se.
notation introduced in Section 3 (c), each entry in the following table symbolizes the nature of the solution to the game when \( i + j = m \), for differing (column) values of \( m \) and differing (row) compositions of \( i + j \), starting in the first row always with \( i = 0 \) and then progressing in the second to \( i = 1 \), etc. Apart from those for \( m = 4 \), the various solutions have already been given in the analysis contained in Sections 1 through 3.

\[
\begin{array}{cccccc}
\text{m} & 1 & 2 & 3 & 4 & 5 \\
\hline
i + j & A_1 & A_2 & A_3 & A_4 & \ldots \\
1 & B_2 & B_1 & B_3 & B_4 & \ldots \\
2 & C_3 & B_2 & A_1 & C_4 & \ldots \\
3 & & C_3 & B_2 & A_1 & \ldots \\
4 & & & D_4 & C_3 & B_2 & A_1 \\
5 & & & & \ldots & \ldots & \ldots \\
\end{array}
\]

The recursive structure of this table is apparent and could clearly be continued indefinitely, whatever \( m \) may be. The rationales for the only new solutions depicted here, those for \( m = 4 \), are each fully evident from the previous discussion. The diagonal entry in the fourth row, where \( i = 3 \) and \( j = 1 \), introduces a new Type D4 of Smith’s reasoning abilities. Its tedious verbal construction requires every prisoner to presume the universal use of \( CK^3 \), \( CK^2 \) and \( CK^1 \) before rounds three, two and one, respectively; and so on.

For a similarly heuristic proof of (c), note that in all the sub-cases discussed in Sections 2 and 3 for which a solution exists, it was discovered at a round whose number can be expressed by the formula \( m - j + 1 \), where in accordance with the discussion in 0 (a) it is understood that in the case \( m = 0 \) the variable \( j \) is to be omitted from the formula; the same formula clearly also fits the solutions for \( m = 4 \).

5. PUBLIC KNOWLEDGE

Suppose that at the end of his address to the prisoners and after the blindfolds have been removed, the King had made the following pronouncement:

‘I may be cruel but I neither lie nor cheat. You have my word that there always will be at least one red hat, so that at least one of you will always go free if he guesses right.'

Provided that none of the prisoners is blind or deaf and that they all believe the king, the commonality of this meagre news is arbitrarily high. Every prisoner heard it, every prisoner is aware that every prisoner heard it; \ldots and so on ad infinitum. Thus this is a case of unbounded commonality, which may be denoted appropriately by \( CK^m \). Such cases are so extreme (although not necessarily uncommon) that like the case of private knowledge they deserve a special name. Call them examples of public knowledge \( CK^m \). They are polar to the case of private knowledge, \( CK^0 \).

Assuming the Theorem to be true it immediately implies that any news, no matter how meagre, is enough to solve the game provided only that it is public knowledge. For then \( i = \infty \), so no matter how large \( m \) may be the condition in part (b) is satisfied. Moreover, since always \( n \geq m \), from part (c) every prisoner would discover his true situation at the very latest by round \((n - 1 + 1) = n \). So with such an announcement the game would always be solved in finite time. The King would be foiled yet again, this time not by his own cheating but by his own meagre sense of fairness.