Reasoning about Knowledge

Ronald Fagin
Joseph Y. Halpern
Yoram Moses
Moshe Y. Vardi

The MIT Press
Cambridge, Massachusetts
London, England

First MIT Press paperback edition, 2003

This book was set in Times Roman and MathTime by Windfall Software (using LaTeX) and was printed and bound in the United States of America.

Library of Congress Cataloging-in-Publication Data

Reasoning about knowledge / Ronald Fagin . . . [et al.].
p. cm.
Includes bibliographical references and index.
ISBN 978-0-262-06162-9 (hc.: alk. paper), 978-0-262-56200-3 (pb.:alk. paper)
1. Knowledge, Theory of. 2. Agent (Philosophy) 3. Reasoning.
I. Fagin, Ronald.
BD181.R38 1995
153.4′—dc20 94-36477
CIP

10 9 8 7 6 5 4

1.1 The "Muddy Children" Puzzle

Reasoning about the knowledge of a group can involve subtle distinctions between a number of states of knowledge. A good example of the subtleties that can arise is given by the "muddy children" puzzle, which is a variant of the well known "wise men" or "cheating wives" puzzles.

> Imagine n children playing together. The mother of these children has told them that if they get dirty there will be severe consequences. So, of course, each child wants to keep clean, but each would love to see the others get dirty. Now it happens during their play that some of the children, say k of them, get mud on their foreheads. Each can see the mud on others but not on his own forehead. So, of course, no one says a thing. Along comes the father, who says, "At least one of you has mud on your forehead," thus expressing a fact known to each of them before he spoke (if $k > 1$). The father then asks the following question, over and over: "Does any of you know whether you have mud on your own forehead?" Assuming that all the children are perceptive, intelligent, truthful, and that they answer simultaneously, what will happen?
>
> There is a "proof" that the first $k - 1$ times he asks the question, they will all say "No," but then the k^{th} time the children with muddy foreheads will all answer "Yes."
>
> The "proof" is by induction on k. For $k = 1$ the result is obvious: the one child with a muddy forehead sees that no one else is muddy. Since he knows that there is at least one child with a muddy forehead, he concludes that he must be the one. Now suppose $k = 2$. So there are just two muddy children, a and b. Each answers "No" the first time, because of the mud on the other. But, when b says "No," a realizes that he must be muddy, for otherwise b would have known the mud was on his forehead and answered "Yes" the first time. Thus a answers "Yes" the second time. But b goes through the same reasoning. Now suppose $k = 3$; so there are three muddy children, a, b, c. Child a argues as follows. Assume that I do not have mud on my forehead. Then, by the $k = 2$ case, both b and c will answer "Yes" the second time. When they do not, he realizes that the assumption was false, that he is muddy, and so will answer "Yes" on the third question. Similarly for b and c.
>
> The argument in the general case proceeds along identical lines.

Let us denote the fact "at least one child has a muddy forehead" by p. Notice that if $k > 1$, that is, more than one child has a muddy forehead, then every child can see at least one muddy forehead, and the children initially all know p. Thus, it would seem that the father does not provide the children with any new information, and so he should not need to tell them that p holds when $k > 1$. But this is false! In fact, as we now show, if the father does not announce p, the muddy children are never able to conclude that their foreheads are muddy.

Here is a sketch of the proof: We prove by induction on q that, no matter what the situation is, that is, no matter how many children have a muddy forehead, all the children answer "No" to the father's first q questions. Clearly, no matter which children have mud on their foreheads, all the children answer "No" to the father's first question, since a child cannot tell apart a situation where he has mud on his forehead from one that is identical in all respects except that he does not have a muddy forehead. The inductive step is similar: By the inductive hypothesis, the children answer "No" to the father's first q questions. Thus, when the father asks his question for the $(q+1)^{\text{st}}$ time, child i still cannot tell apart a situation where he has mud on his forehead from one that is identical in all respects except that he does not have a muddy forehead, since by the induction hypothesis, the children will answer "No" to the father's first q questions whether or not child i has a muddy forehead. Thus, again, he does not know whether his own forehead is muddy.

So, by announcing something that the children all know, the father somehow manages to give the children useful information! How can this be? Exactly what *is* the role of the father's statement? Of course, the father's statement did enable us to do the base case of the induction in the proof, but this does not seem to be a terribly satisfactory answer. It certainly does not explain what information the children gained as a result of the father's statement.

We can answer these questions by using the notion of common knowledge described in the previous section. Let us consider the case of two muddy children in more detail. It is certainly true that before the father speaks, everyone knows p. But it is not the case that everyone knows that everyone knows p. If Alice and Bob are the only children with muddy foreheads, then before the father speaks, Alice considers it possible that she does not have mud on her forehead, in which case Bob does not see anyone with a muddy forehead and so does not know p. After the father speaks, Alice does know that Bob knows p. After Bob answers "No" to the father's first question, Alice uses her knowledge of the fact that Bob knows p to deduce that her

own forehead is muddy. (Note that if Bob did not know p, then Bob would have said "No" the first time even if Alice's forehead were clean.)

We have just seen that if there are only two muddy children, then it is not the case that everyone knows that everyone knows p before the father speaks. However, if there are three muddy children, then it *is* the case that everyone knows that everyone knows p before the father speaks. If Alice, Bob, and Charlie have muddy foreheads, then Alice knows that Bob can see Charlie's muddy forehead, Bob knows that Charlie can see Alice's muddy forehead, etc. It is not the case, however, that everyone knows that everyone knows that everyone knows p before the father speaks. In general, if we let $E^k p$ represent the fact that everyone knows that everyone knows ... (k times) p, and let Cp represent the fact that p is common knowledge, then we leave it to the reader to check that if exactly k children have muddy foreheads, then $E^{k-1} p$ holds before the father speaks, but $E^k p$ does not. It turns out that when there are k muddy children, $E^k p$ suffices to ensure that the children with muddy foreheads will be able to figure it out, while $E^{k-1} p$ does not. The father's statement actually converts the children's state of knowledge from $E^{k-1} p$ to Cp. With this extra knowledge, they can deduce whether their foreheads are muddy.

The careful reader will have noticed that we made a number of implicit assumptions in the preceding discussion over and above the assumption made in the story that "the children are perceptive, intelligent, and truthful." Suppose again that Alice and Bob are the only children with muddy foreheads. It is crucial that both Alice and Bob *know* that the children are intelligent, perceptive, and truthful. For example, if Alice does not know that Bob is telling the truth when he answers "No" to the father's first question, then she cannot answer "Yes" to the second question (even if Bob is in fact telling the truth). Similarly, Bob must know that Alice is telling the truth. Besides its being known that each child is intelligent, perceptive, and truthful, we must also assume that each child knows that the others can see, that they all hear the father, that the father is truthful, and that the children can do all the deductions necessary to answer the father's questions.

Actually, even stronger assumptions need to be made. If there are k children with muddy foreheads, it must be the case that everyone knows that everyone knows ... ($k - 1$ times) that the children all have the appropriate attributes (they are perceptive, intelligent, all hear the father, etc.). For example, if there are three muddy children and Alice considers it possible that Bob considers it possible that Charlie might not have heard the father's statement, then she cannot say "Yes" to the father's third question (even if Charlie in fact did hear the father's statement and Bob

knows this). In fact, it seems reasonable to assume that all these attributes are common knowledge, and, indeed, this assumption seems to be made by most people on hearing the story.

To summarize, it seems that the role of the father's statement was to give the children common knowledge of p (the fact that at least one child has a muddy forehead), but the reasoning done by the children assumes that a great deal of common knowledge already existed in the group. How does this common knowledge arise? Even if we ignore the problem of how facts like "all the children can see" and "all the children are truthful" become common knowledge, there is still the issue of how the father's statement makes p common knowledge.

Note that it is not quite correct to say that p becomes common knowledge because all the children hear the father. Suppose that the father had taken each child aside individually (without the others noticing) and said "At least one of you has mud on your forehead." The children would probably have thought it a bit strange for him to be telling them a fact that they already knew. It is easy to see that p would not become common knowledge in this setting.

Given this example, one might think that the common knowledge arose because all the children *knew* that they all heard the father. Even this is not enough. To see this, suppose the children do not trust each other, and each child has secretly placed a miniature microphone on all the other children. (Imagine that the children spent the previous summer at a CIA training camp.) Again the father takes each child aside individually and says "At least one of you has a muddy forehead." In this case, thanks to the hidden microphones, all the children know that each child has heard the father, but they still do not have common knowledge.

A little more reflection might convince the reader that the common knowledge arose here because of the *public* nature of the father's announcement. Roughly speaking, the father's public announcement of p puts the children in a special situation, one with the property that all the children know both that p is true and that they are in this situation. We shall show that under such circumstances p is common knowledge. Note that the common knowledge does not arise because the children somehow deduce each of the facts $E^k p$ one by one. (If this were the case, then arguably it would take an infinite amount of time to attain common knowledge.) Rather, the common knowledge arises all at once, as a result of the children being in such a special situation. We return to this point in later chapters.

Exercises

1.1 The *aces and eights* game is a simple game that involves some sophisticated reasoning about knowledge. It is played with a deck consisting of just four aces and four eights. There are three players. Six cards are dealt out, two to each player. The remaining two cards are left face down. Without looking at the cards, each of the players raises them up to his or her forehead, so that the other two players can see them but he or she cannot. Then all of the players take turns trying to determine which cards they're holding (they do not have to name the suits). If a player does not know which cards he or she is holding, the player must say so. Suppose that Alice, Bob, and you are playing the game. Of course, it is common knowledge that none of you would ever lie, and that you are all perfect reasoners.

(a) In the first game, Alice, who goes first, holds two aces, and Bob, who goes second, holds two eights. Both Alice and Bob say that they cannot determine what cards they are holding. What cards are you holding? (Hint: consider what would have happened if you held two aces or two eights.)

(b) In the second game, you go first. Alice, who goes second, holds two eights. Bob, who goes third, holds an ace and an eight. No one is able to determine what he or she holds at his or her first turn. What do you hold? (Hint: by using part (a), consider what would have happened if you held two aces.)

(c) In the third game, you go second. Alice, who goes first, holds an ace and an eight. Bob, who goes third, also holds an ace and an eight. No one is able to

determine what he or she holds at his or her first turn; Alice cannot determine her cards at her second turn either. What do you hold?

*** 1.2** Show that in the aces and eights game of Exercise 1.1, someone will always be able to determine what cards he or she holds. Then show that there exists a situation where only one of the players will be able to determine what cards he or she holds, and the other two will never be able to determine what cards they hold, no matter how many rounds are played.

1.3 The *wise men puzzle* is a well-known variant of the muddy children puzzle. The standard version of the story goes as follows: There are three wise men. It is common knowledge that there are three red hats and two white hats. The king puts a hat on the head of each of the three wise men, and asks them (sequentially) if they know the color of the hat on their head. The first wise man says that he does not know; the second wise man says that he does not know; then the third wise man says that he knows.

(a) What color is the third wise man's hat?

(b) We have implicitly assumed in the story that the wise men can all see. Suppose we assume instead that the third wise man is blind and that it is common knowledge that the first two wise men can see. Can the third wise man still figure out the color of his hat?

Notes

The idea of a formal logical analysis of reasoning about knowledge seems to have first been raised by von Wright [1951]. As we mentioned in the text, Hintikka [1962] gave the first book-length treatment of epistemic logic. Lenzen [1978] gives an overview of the work in epistemic logic done in the 1960's and 1970's. He brings out the arguments for and against various axioms of knowledge. The most famous of these arguments is due to Gettier [1963], who argued against the classical interpretation of knowledge as true, justified belief; his work inspired many others. Gettier's arguments and some of the subsequent papers are discussed in detail by Lenzen [1978]. For recent reviews of the subject, see the works by Halpern [1986, 1987,

1995], by Meyer, van der Hoek, and Vreeswijk [1991a, 1991b] (see also [Meyer and Hoek 1995]), by Moses [1992], and by Parikh [1990].

As we mentioned, the original work on common knowledge was done by Lewis [1969] in the context of studying conventions. Although McCarthy's notion of what "any fool" knows goes back to roughly 1970, it first appears in a published paper in [McCarthy, Sato, Hayashi, and Igarishi 1979]. The notion of knowledge and common knowledge has also been of great interest to economists and game theorists, ever since the seminal paper by Aumann [1976]. Knowledge and common knowledge were first applied to multi-agent systems by Halpern and Moses [1990] and by Lehmann [1984]. The need for common knowledge in understanding a statement such as "What did you think of the movie?" is discussed by Clark and Marshall [1981]; a dissenting view is offered by Perrault and Cohen [1981]. Clark and Marshall also present an example of nested knowledge based on the Watergate scandal, mentioning Dean and Nixon. The notion of distributed knowledge was discussed first, in an informal way, by Hayek [1945], and then, in a more formal way, by Hilpinen [1977]. It was rediscovered and popularized by Halpern and Moses [1990]. They initially called it *implicit knowledge*, and the term "distributed knowledge" was suggested by Jan Pachl.

The muddy children puzzle is a variant of the "unfaithful wives" puzzle discussed by Littlewood [1953] and Gamow and Stern [1958]. Gardner [1984] also presents a variant of the puzzle, and a number of variants of the puzzle are discussed by Moses, Dolev, and Halpern [1986]. The version given here is taken almost verbatim from [Barwise 1981]. The aces and eights game in Exercise 1.1 is taken from [Carver 1989]. Another related puzzle is the so-called "Conway paradox", which was first discussed by Conway, Paterson, and Moscow [1977], and later by Gardner [1977]. It was analyzed in an epistemic framework by van Emde Boas, Groenendijk, and Stokhof [1980]. An extension of this puzzle was considered by Parikh [1992]. The wise men puzzle discussed in Exercise 1.3 seems to have been first discussed formally by McCarthy [1978], although it is undoubtedly much older. The well-known *surprise test paradox*, also known as the *surprise examination paradox*, the *hangman's paradox*, or the *unexpected hanging paradox*, is quite different from the wise men puzzle, but it too can be analyzed in terms of knowledge. Binkley [1968] does an analysis that explicitly uses knowledge; Chow [1998] gives a more up-to-date discussion. Halpern and Moses [1986] give a slightly different logic-based analysis, as well as pointers to the literature.