Chapter 1

IT'S COMMON KNOWLEDGE

> Mathematics is the science which uses easy words for hard ideas.
>
> Edward Kasner and James Newman

Part of Stevie Nicks's 2003 lyrics of the Fleetwood Mac song 'Everybody Finds Out' might describe the reader's reaction to the subject matter of this first chapter:

> I know you don't agree…
> Well, I know you don't agree.

The song's title also finds its way into the title of an episode of the NBC sitcom television series *Friends*, 'The One Where Everybody Finds Out', which was first aired in February 1999 and which contains the following dialogue:

> Rachel: Phoebe just found out about Monica and Chandler.
> Joey: You mean how they're friends and nothing more? [Glares at Rachel]
> Rachel: No. Joey, she knows! We were at Ugly Naked Guy's apartment and we saw them doing it through the window. [Joey gasps] Actually, we saw them doing it up against the window.
> Phoebe: Okay, so now they know that you know and they don't know that Rachel knows?

Joey: Yes, but y'know what? It doesn't matter who knows what. Now, enough of us know that we can just tell them that we know! Then all the lying and the secrets would finally be over!
Phoebe: Or, we could not tell them we know and have a little fun of our own.
Rachel: Wh-what do you mean?

We will consider what brings together pop lyrics, popular television and an important idea which purveys a considerable body of mathematics and its applications.

Common and Mutual Knowledge

It seems unlikely that the above discourse was meant to plumb the depths of mathematical logic but it does draw an important distinction between two superficially equivalent concepts: *mutual knowledge* and *common knowledge*. We might, for example, suggest that in everyday language it is common knowledge that the capital city of Australia is Canberra and mean that all people who know of the country will reasonably be aware of that fact. As another example, we might say that it is common knowledge that all road users know that a red traffic light means 'stop' and a green traffic light 'go'. The usage of an ordinary expression such as 'common knowledge' is fine in normal circumstances but we are going to deal with a stricter interpretation of the term and distinguish it from its cousin, mutual knowledge.

Mathematics is given to using ordinary words for technical purposes: *group*, *ring*, *field*, *rational*, *transcendental*, etc., each have their standard dictionary definitions yet each of them means something entirely different, precise and technical within the mathematical world. The same is true of the phrase *common knowledge*, the everyday use of which suggests that what is being referred to is known to all. The crucial point is that it would not matter very much whether or not an individual knows that another knows that the capital city of Australia is Canberra, but in order to ensure safe traffic flow it is not sufficient that

all road users are aware of the colour convention used with traffic lights; it must be the case that they know that all other road users are aware of the convention, otherwise a driver might see a car approaching a red light and wonder whether or not its driver is aware of the convention to stop there.

So, we make two definitions. The first is that of *mutual knowledge*. A statement S is said to be *mutual knowledge* among a group of people if each person in that group knows S. Mutual knowledge by itself implies nothing about what, if any, knowledge anyone attributes to anyone else. It is sufficient that the Canberra example is one of mutual knowledge. The technical definition of *common knowledge* brings about a deeper implication: that everyone knows that everyone knows (and everyone knows that everyone knows that everyone knows, and so on) S. The traffic light example requires common knowledge.

In the dialogue above, Phoebe's statement,

> Okay, so now they know that you know and they don't know that Rachel knows?

distinguishes between the common knowledge shared between Joey and the couple Monica and Chandler and the mutual knowledge shared between Rachel and them.

It is possible to convert mutual into common knowledge. For example, we could assemble a group of strangers in a room and then make the statement: the capital city of Australia is Canberra. If we assume that each individual already knew the fact (and therefore it was already mutual knowledge), at first glance the announcement seems to add nothing, but it has transformed mutual knowledge into common knowledge, with everybody in the room now knowing that everybody in the room knows that the capital city of Australia is Canberra. It is this feature that is central to the main conundrum of the chapter.

There is a well-known example of the phenomenon in children's literature. In Hans Christian Andersen's fable *The Emperor's New Clothes*, two scoundrels convince the vain emperor that they could make a magnificent cloth of silk and gold threads which would be 'invisible to everyone who was stupid or not fit

for his post'. After the emperor gave them money and materials to make the royal garments, they dressed him in nothing at all. Not even the emperor, much less his courtiers, dared admit to not seeing any clothes for fear of being branded stupid or incompetent. A ceremonial parade was arranged in order to display the wondrous new clothes and the public applauded as the emperor passed by.

> All the people standing by and at the windows cheered and cried, 'Oh, how splendid are the emperor's new clothes.'

Then a child commented,

> But he hasn't got anything on.

From that moment, what had been the mutual knowledge that the emperor was naked became common knowledge.

This is more than semantic pedantry and we will consider an infamous example of the implication of converting mutual to common knowledge. The technique of mathematical induction will be used and this is reviewed in the appendix (page 221).

A Case of Red and Blue Hats

Suppose that a group of people is assembled in a room and also a number of hats, one for each, coloured either red or blue (accepting that all could be of one colour). For definiteness we will suppose that exactly fifteen of the hats are red, which means that the remainder are blue, although the participants are not aware of this distribution. We will also suppose that each individual is a perfect logician.

A hat is placed on each person's head in such a way that its colour is unknown to that individual but is seen by everyone else. The group of people then sits in the room looking at each other, without communication, and with a clock, which strikes every hour on the hour, available for all to see and hear. Each is instructed to leave the room immediately after the clock strike after which they are certain that they are wearing a red hat.

The group will simply sit in the room, waiting as the clock strikes hour after hour. Those wearing a red hat will see fourteen

red hats and those wearing a blue hat will see fifteen red hats; with no extra information, none of them can be certain of the colour of the hat they are wearing: are there fourteen, fifteen or sixteen red hats? Fortunately, a visitor arrives in the room, looks around at the hats being worn and announces, 'At least one person here is wearing a red hat.'

This hardly seems revelatory. Notwithstanding the seeming irrelevance of the announcement, once it is made it is certain that after the subsequent fifteenth strike of the clock, all fifteen people who are wearing red hats will simultaneously walk out of the room.

To consider the reasoning it will be convenient to adopt some notation. Represent the statement 'at least one person is wearing a red hat' by the symbol R_1 and the statement 'A knows X' by the expression $A \rightarrow X$.

First, consider the case of one red hat. Before the statement the wearer, A, sees all blue hats and can have no idea of the colour of his own hat; that is, $A \not\rightarrow R_1$. After the announcement $A \rightarrow R_1$ and he will be certain that his hat is red and will walk out after the next clock strike, the first after the announcement. The information that was conveyed by the announcement results in an immediate resolution of the situation.

Now we will deal with two red hats. Before the announcement, R_1 is mutual but not common knowledge. That is, everyone can see at least one red hat and, if the wearers of the red hats are A and B, then $A \rightarrow R_1$ and $B \rightarrow R_1$, since each can see the other's red hat. Yet, $A \not\rightarrow (B \rightarrow R_1)$, since $B \rightarrow R_1$ is a direct result of B seeing A's red hat and A has no idea whether or not his hat is indeed red. The announcement tells everybody that R_1 is true and so it is now the case that $A \rightarrow (B \rightarrow R_1)$ (and $B \rightarrow (A \rightarrow R_1)$). Information has been acquired by the announcement; what was mutual knowledge among the red-hat wearers has become common knowledge among them. Now the clock strikes for the first time and none can conclude the colour of their hat: it could be that there is one red hat, in which case, from A's point of view, B is wearing it. Then it strikes a second time and matters change. A argues that, since B did not leave after the first strike

of the clock, it must be that he saw a red hat and therefore that there are two such, one on each of the heads of A and B: both will leave the room.

With three red hats, before the announcement the following typifies the situation for red-hat wearers A, B and C: $B \to R_1$, as B can see red hats on both A and C; $A \to (B \to R_1)$, as A can see a red hat on C, but $C \not\to (A \to (B \to R_1))$, since C has no idea whether his hat is red or blue.

After the announcement, R_1 again becomes common knowledge and so $C \to (A \to (B \to R_1))$ and once again information is contained within the seemingly innocent statement and the same argument as above establishes that all three leave after the third strike of the clock.

The reasoning continues with ever deeper levels of knowledge gained as the number of red hats grows with the announcement causing the strike-out of the first arrow in the knowledge chain to disappear in every case: everybody knows that everybody knows that… there is at least one red hat. From this, it is a matter of waiting in order to exclude all possibilities until in the end only one remains; in our case, that all fifteen red head wearers know that there are precisely fifteen red hats.

The 'reasoning continues' type of argument is one which is normally susceptible to proof by induction and we give one such below.

The induction is taken over the clock strike, with R_i taken to mean 'at least i people are wearing red hats'. Now suppose that at the ith strike of the clock R_i is common knowledge. If no red-hatted individual can tell if his hat is red, it must be that R_{i+1} is true since each must be seeing at least i red hats, otherwise he will be able to tell that his hat is red; this together with his own red hat gives the result. This means that on the fifteenth strike of the clock that there are at least fifteen red hats is common knowledge; but the red-hat wearers can only see fourteen red hats and so they must conclude that their hat is indeed red, and will walk out.

The puzzle is one of many variants—with luminaries such as John Edensor Littlewood giving their names to some of

them—they all reduce to the same fundamental concept and they are all very, very confusing!

The importance of common knowledge extends far and wide in mathematical application, including the fields of economics, game theory, philosophy, artificial intelligence and psychology. Perhaps the concept dates back as far as 1739 when, in his *Treatise of Human Nature*, the Scottish philosopher David Hume argued that, in order to engage in coordinated activity, all participants must know what behaviour to expect from each other. It is not difficult for the modern author to have empathy with Hume when he (too critically) judged the initial public reaction to the work as such that it 'fell dead-born from the press, without reaching such distinction as even to excite a murmur among the zealots.' It is now generally considered to be one of the most important books in the development of modern philosophy.

As a final problem, we will consider a situation reminiscent of the above in that a striking clock counts out seemingly irrelevant time periods but in which a seemingly irrelevant statement is replaced by a seemingly unhelpful condition.

Consecutive Integers

Two people, A and B, are assigned positive integers; secretly, they are each told their integer and also that the two integers are consecutive. The two sit in a room in which there is a clock, which strikes every hour on the hour. They may not communicate in any way, but they are instructed to wait in the room until one knows the other's number and then to announce that number after the strike of the clock following the revelation of that information.

Both seem destined to stay in the room forever. The clock will relentlessly strike the hour with the two participants seemingly waiting for help that never comes: imagine sitting in the room with, for example, the knowledge that your number is 57; you can have no idea whether the other number is 56 or 58—or can you?

In fact, there is a hidden advantage in the clock striking and knowing that the numbers are consecutive, which our intuition

can easily fail to exploit. A careful use of induction can succeed in that exploitation, and having done so should convince us that at some stage one of the two people will leave the room.

To get a feel for what is really happening, suppose that A's number is 1, then it must be that B's number is 2 and after the first strike of the clock A will announce that B has the number 2. Now take the next case and suppose that A's number is 2. This means that B's number is either 1 or 3. If it is 1, B will announce after the first strike of the clock, as above; if the announcement is not made, A will know that B's number is 3 and announce this fact after the second strike of the clock. The argument can be continued methodically and is best done so using induction to give the remarkable result that *the person whose number is n will announce that the other player's number is $n+1$ after the nth strike of the clock.*

In fact, the proof is easy. We have already argued that the statement is true if the lower number is 1. Now let us suppose that the statement is true when the lower number is k and that A is given the number $k+1$. Then if B holds k, by the induction hypothesis he will announce A's number after the kth strike of the clock, otherwise B holds $k+2$ and A will know this to be the case after the kth strike of the clock and so announce B's number after the $(k+1)$th strike of the clock, and the induction is complete.