Note 3 Supplement: Common Knowledge

Computer Science 70
University of California, Berkeley

Summer 2018

In this note, we look at a curious logic puzzle.

> One hundred green-eyed dragons inhabit a remote island. The dragons have a rule: if any dragon discovers that it has green eyes, then it must commit ritual suicide the day after. Despite this unforgiving rule, the dragons live peacefully on the island.
>
> One day, a visitor arrives on the island and says, "I see a dragon with green eyes." The visitor leaves the island.
>
> One hundred days after the visitor arrived on the island, all one hundred dragons commit ritual suicide. Why?

One observation is that if one dragon commits ritual suicide on some day, then every dragon must commit ritual suicide on that day by symmetry (there is nothing distinguishing the green-eyed dragons, so there is no reason for one dragon to commit ritual suicide first).

The puzzling aspect of the outcome is that the visitor does not seem to add any new information. Since every dragon sees that every other dragon has green eyes, then each dragon already knew the truth of the visitor's statement, even before the visitor said it.

To understand this puzzle, consider first the situation when the island only consists of two green-eyed dragons. Since each dragon can see that the other dragon has green eyes, each dragon knows the truth of the statement

$$\text{"There is at least one green-eyed dragon on the island."} \qquad (\star)$$

However, does each dragon on the island know that the *other* dragon knows $(\star)$? The answer is *no*, because each dragon must leave open the possibility

that its own eye color is either green, or not green. If the dragon's eye color is green, then the other dragon would indeed know ($\star$). If the dragon's eye color is not green, then the other dragon would not know ($\star$) since the other dragon cannot see its own eye color.

The visitor has thus changed the situation on the island so that every dragon knows the truth of the statement

$$\text{"Every dragon knows } (\star)\text{."} \qquad (\star\star)$$

The truth of ($\star\star$) comes because every dragon has witnessed every other dragon hearing the visitor's statement. However, since every dragon has witnessed every other dragon witnessing the visitor's statement, then every dragon knows the truth of the statement:

$$\text{"Every dragon knows } (\star\star)\text{."} \qquad (\star\star\star)$$

We can continue this reasoning, which brings us to a concept called **common knowledge**. Common knowledge refers to a fact which is known by every individual; but also every individual knows that every individual knows this fact; and also that every individual knows that every individual knows that every individual knows this fact, and so forth.

Returning to the case of two dragons, on the first day, no dragon has enough information to conclude that it has green eyes, despite it being common knowledge that there exists a dragon with green eyes on the island. So, both dragons survive until the second day, upon which each dragon reasons that if it did *not* have green eyes, then the other dragon would have had enough information on the first day to commit ritual suicide, and observing that the other dragon survived until the second day, concludes that both dragons have green eyes. Thus *both* dragons commit ritual suicide on the second day.

The situation for two dragons can be thought of as a chain of reasoning which is two levels deep, and the situation for one hundred dragons requires reasoning which is nested one hundred levels deep. It is unreasonable to keep track of one hundred levels of reasoning in your head at once, but induction allows us to make sense of the situation.

We will do one more case to get a feeling for the situation: suppose there are three green-eyed dragons on the island, which we name A, B, and C. On the first day, Dragon A does not know its own eye color, so it must leave open the possibility that it has blue eyes. It then wonders what Dragon B is thinking; since Dragon B must leave open the possibility that it has blue

eyes, then Dragon B might think that Dragon C sees two blue-eyed dragons. In other words, Dragon A thinks that Dragon B might think that Dragon C might think that Dragons A and B have blue eyes. In such a situation, upon being told ($\star$), Dragon C would then conclude that it must be the only dragon with green eyes, and it would kill itself on the first day.

On the second day, Dragon C is still alive. So, Dragon A thinks that if it had blue eyes, then Dragon B, upon realizing that Dragon C is still alive, would conclude that it has green eyes (otherwise Dragon C would be dead), so Dragon B would kill itself on the second day.

On the third day, Dragon B is still alive. So, Dragon A concludes all three dragons must have green eyes, and Dragon A kills itself on the third day. Since each of the three dragons follows the same reasoning, then all dragons kill themselves on the third day.

Theorem 1. *For every positive integer n, if there are n green-eyed dragons on the island, then n days after the visitor arrives on the island, every dragon will commit ritual suicide.*

Proof. The base case is easy to verify. Inductively, for a positive integer n, assume that if there are n green-eyed dragons on the island, then they will commit suicide n days after the visitor's arrival. Now, consider the situation when there are $n+1$ green-eyed dragons on the island.

In the first n days, nothing will happen, as no dragon has enough information to conclude that it has green eyes. On day $n+1$, however, the dragon sees that the other dragons have not committed ritual suicide by day n, which means there must be at least $n+1$ green-eyed dragons. The dragon therefore concludes that it has green eyes and commits ritual suicide. □