

- 25 (the unexpected egg) There are two boxes; one is red and the other is blue. One box has an egg in it; the other is empty. You are to look first in the red box, then if necessary in the blue box, to find the egg. But you will not know which box the egg is in until you open the box and see the egg. You reason as follows: “If I look in the red box and find it empty, I’ll know that the egg is in the blue box without opening it. But I was told that I would not know which box the egg is in until I open the box and see the egg. So it can’t be in the blue box. Now I know it must be in the red box without opening the red box. But again, that’s ruled out, so it isn’t in either box.”. Having ruled out both boxes, you open them and find the egg in one unexpectedly, as originally stated. Formalize the given statements and the reasoning, and thus explain the paradox.

After trying the question, scroll down to the solution.

§ Let  $r$  mean “the egg is in the red box” and let  $b$  mean “the egg is in the blue box”. Let  $Kr$  mean “you know the egg is in the red box before seeing it there” and let  $Kb$  mean “you know the egg is in the blue box before seeing it there”. We are told (as axioms)

- (a)  $r \neq b$
- (b)  $\neg Kr \wedge \neg Kb$

Perhaps there is one other axiom that is implicit in the use of the word “know”:

- (c)  $(Kr \Rightarrow r) \wedge (Kb \Rightarrow b)$

If something is true, you might not know it; but if you know it, then it is indeed true. This axiom is certainly debatable, and we don't need it because it follows directly from (b) anyway. So far, we have consistent axioms but we can't prove which box the egg is in. The reasoner (the “you” in the question) begins with the statement

- (d)  $\neg r \Rightarrow Kb$

This statement does not follow from the three axioms (a), (b), and (c); suppose  $b$  is  $\top$  and the other three variables are all  $\perp$ ; then all axioms are satisfied but (d) is not. The reasoner must be assuming one more axiom, perhaps

- (e)  $(r \Rightarrow Kr) \wedge (b \Rightarrow Kb)$

In words: if  $r$  is true, then I know  $r$ , and if  $b$  is true, then I know  $b$ . Now (d) does follow like this:

$$\begin{array}{ll} \neg r & \text{use (a)} \\ = b & \text{use the second part of (e)} \\ \Rightarrow Kb & \end{array}$$

The reasoner now invokes axiom (b), specifically the second conjunct, to contradict the consequent of (d), and so to contradict its antecedent, and correctly concludes  $r$ , and from (a) concludes  $\neg b$  (“it can't be in the blue box”). Now the first part of (e) is needed to get to the reasoner's next statement, which is  $Kr$ . Again (b) is invoked, this time the first conjunct, to contradict  $Kr$  (“But again, that's ruled out”). The reasoner now concludes  $\neg r \wedge \neg b$ , though it is not clear how. Nonetheless it is a correct conclusion, as would any other be, for the axioms (a), (b), and (e) are inconsistent. Inconsistent information is the usual explanation of paradoxes. But how could the reasoner have made the mistake of assuming (e)? I'll explain in a moment.

This puzzle is a simpler version of one in a book by Martin Gardner: *the Unexpected Hanging*, Simon&Schuster, 1969. A judge tells a prisoner that he will be hanged on one of the days of the coming week, but he won't know which one until it arrives. The prisoner comes to the conclusion that he will not be hanged, and then he is hanged, unexpectedly. The egg puzzle is identical except that 7 has been reduced to 2. Reducing 7 to 2 shortens a chain of reasoning, but does not change the character of the puzzle at all. And we can further reduce 2 to 1. There is one box, and in the box there is an egg. It's there all right; that's part of the given information. So far, so good, but the next piece of given information is strange: you will not know the egg is in the box until you open the box and see it. First you are told something, then in the next breath you are told that you don't know what you were just told. To untangle the mess, let me suppose there are three people. The first is the narrator, who tells the story, and gives the given information. The second person is us, to whom the narrator is speaking. We have to believe the narrator. We cannot distinguish what we know from what we are told by the narrator. The third person is the person the narrator is talking about, who looks in the box. The narrator whispers to us that there is an egg in the box ( $a$ ), and that the looker doesn't know it ( $\neg Ka$ ). We might also (unnecessarily) assume  $Ka \Rightarrow a$ , which says that if the looker did know it, it would be there. But we are not even tempted to assume  $a \Rightarrow Ka$ , which says that if it is there the looker knows it. When the looker looks in the box, the egg will be unexpected. We know what the narrator tells us, but the looker

doesn't. The way the puzzle was presented, we and the looker were identified as one person, and that creates the problem.