One hundred prisoners and a lightbulb – for ESSLLI 2008

Hans van Ditmarsch, Computer Science, University of Otago, New Zealand
hans@cs.otago.ac.nz

August 19, 2008

The following is known as the 'one hundred prisoners and a lightbulb' problem.

> *A group of 100 prisoners, all together in the prison dining area, are told that they will be all put in isolation cells and then will be interrogated one by one in a room containing a light with an on/off switch. The prisoners may communicate with one another by toggling the light-switch (and that the only way in which they can communicate). The light is initially switched off. There is no fixed order of interrogation, or interval between interrogations, and the same prisoner may be interrogated again at any stage. When interrogated, a prisoner can either do nothing, or toggle the light-switch, or announce that all prisoners have been interrogated. If that announcement is true, the prisoners will (all) be set free, but if it is false, they will all be executed. While still in the dining room, and before the prisoners go to their isolation cells (forever), can the prisoners agree on a protocol that will set them free (assuming that at any stage every prisoner will be interrogated again sometime)?*

The riddle appears to have been around for at least five years. We made some investigation on the puzzle's origin. Our source was the ESSLLI 2003 logic summerschool in Vienna, where it was—apparently incorrectly—said to originate with Moshe Vardi. We did not find informal references before 2001. William Wu (see `http://www.ocf.berkeley.edu/~wwu/papers/100prisonersLightBulb.pdf`) mentions hearing about the riddle in 2001 and cites an IBM Research site `http://domino.watson.ibm.com/Comm/wwwr_ponder.nsf/challenges/July2002.html` where it is mentioned (in a 23 prisoner version) "This puzzle has been making the rounds of Hungarian mathematicians' parties." We did not find formal references. We thank Moshe Vardi for his advice.

Of course, the answer to the riddle is: "Yes, they can." We solve the riddle for an arbitrary number of n prisoners. For $n = 1$ and $n = 2$ it is trivial. For $n > 2$ a protocol is as follows:

> The n prisoners appoint one amongst them as the 'counter'. All non-counting prisoners follow the following protocol: the first time they enter the room when the light is off, they turn it on; on all other occasions, they do nothing. The counter follows a different protocol. The first $n-2$ times that the light is on when he enters the interrogation room, he turns it off. Then the next time he enters the room when the light is on, he (truthfully) announces that everybody has been interrogated.[1]

[1] The riddle can also be solved when it is not known if the light is initially on or off. In that case the solution is for the non-counting prisoners to turn light on *twice* only if it is off, and the counter announces that everybody has been interrogated after he has turned off the light $2n - 2$ times.

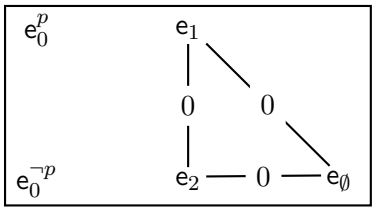

Figure 1: The update model for an interrogation for the situation with three prisoners. See accompanying text for explanations.

In order to model this as a dynamic multi-agent system, we need to provide an initial epistemic model and the updates that are possible in that model. The n prisoners are named $0, \ldots, n-1$. Prisoner 0 is the counter. In the protocol above we are only interested in the information the counter has. It is therefore sufficient only to model his information, and we can model this as a single-agent system. Atomic proposition p stands for for 'the light is on', and atomic propositions q_i, for $1 < i \leq n-1$, for 'prisoner i has turned on the light'. Formula $\bigwedge_{1<i\leq n-1} q_i$ is true when all prisoners except the counter have been interrogated.

We first define the update model that describes an event that a prisoner is interrogated. The ordinary prisoners execute the protocol that they turn on the light, if it is off and if the prisoner has not turned it on before. This event e_i can be seen as the simultaneous assignment.

$$p := q_i \to p \text{ and } \; q_i := p \to q_i$$

The counter executes a different protocol. He never turns on the light. He only turns it off if it is on, otherwise he leaves it alone. This can be modelled as event e_0^p consisting of assignment

$$p := \bot$$

and event $\mathsf{e}_0^{\neg p}$ consisting of precondition $\neg p$ only. In the single agent setting it is meaningless to model the announcement of the counter, and it is sufficient to show that eventually the precondition of that announcement is satisfied, namely that he eventually knows that all the other prisoners have turned on the light once. Finally, event $\mathsf{e}_\emptyset$ models that nothing happens when alternatively a prisoner could have been interrogated; it is therefore a 'skip' action. An overview of the different events is

$\mathsf{e}_\emptyset$	skip	'skip' means 'if $\top$ then $\emptyset$'
e_i	$p := q_i \to p$ and $\; q_i := p \to q_i$	for each $i > 0$
e_0^p	if $\; p$ then $\; p := \bot$	
$\mathsf{e}_0^{\neg p}$	$\neg p$	

Agent 0 cannot distinguish between events e_i and $\mathsf{e}_\emptyset$ but e_0^p and $\mathsf{e}_0^{\neg p}$ that involve himself are of course distinguishable from those and from one another. The update model for the case of three prisoners is given is Figure 1.

We proceed with describing the epistemic model for the problem. The states in the model for the problem are characterized by the facts p and q_i true there. Although we defined the execution of an event as a transition from one epistemic state to another epistemic state, a different perspective is to see this as a 'shift' between points in a single (larger) epistemic model, i.e., more as a 'run' through a system. This allows for a simpler visualization. For the case of three prisoners, see Figure 2. Let us take a close look at this picture.

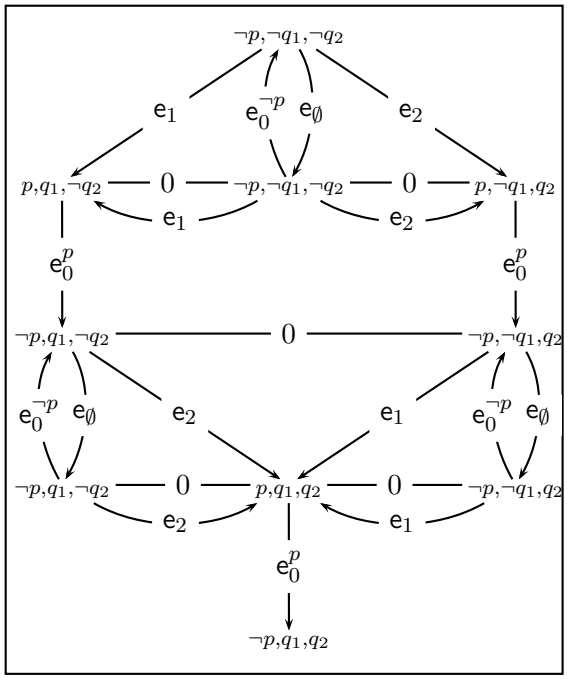

Figure 2: The initial epistemic state and the results of executing the update model for all possible events are pictured for the case of *three* prisoners trying to escape. The states are indicated by an atomic description. We assume reflexivity and transitivity of access for agent 0. We also did not draw the reflexive arrows for events that can be executed but do not yield information change. For example, in the top-left state events e_1, e_2, and $\mathsf{e}_\emptyset$ can also be executed but have no effect. In the bottom state agent 0 knows that all other prisoners have turned on the light at least once.

Initially all prisoners are in the dining area so there is still common knowledge that nobody has been interrogated: a singleton model with state $(\neg p, \neg q_1, \neg q_2)$ models that. This is the top state in the picture. The event that the prisoners leave the dining area is modelled as the non-deterministic execution of four possible events, namely all events with precondition $\neg p$: e_1, e_2, $\mathsf{e}_0^{\neg p}$, and $\mathsf{e}_\emptyset$. The last represents that even when noone has been interrogated yet, the counter considers it possible that someone has been interrogated. That this action is necessary becomes clear when the counter is interrogated himself and finds the light still off. He will then learn that noone was interrogated yet: a transition back to the top state of the model. Without the $\mathsf{e}_\emptyset$-event we could not have modelled this. When one of the other prisoners is interrogated, this leads to the light being on. No information change occurs as long as the counter is not interrogated. When this happens, he turns the light off and learns that at least one prisoner has been interrogated. The counter considers it possible that prisoner 1 has turned on the light and that prisoner 2 has turned on the light. Upon leaving the interrogation room, again the counter immediately considers it possible that other prisoners may have been interrogated, modelled by another informative $\mathsf{e}_\emptyset$ transition. This possibility can again be excluded by observing that the light is still off during yet another interrogation. Eventually the other prisoner is also interrogated, which leads to the light being on again and once this is observed, the counter knows that all the other prisoners have turned on the light once. Here we model this by another e_0^p-transition, now to the bottom state of the picture. The counter now knows that everyone has been interrogated (and in principle he can make an announcement to that effect).

It would be nice to also investigate what happens if we do take the information of the non-counting prisoners into account. Suppose for example that one such agent is always interrogated before the counter. That would mean the prisoner gets the same information as the counter, just a bit quicker. Indeed, this prisoner could announce that everyone has been interrogated before the counter. Then we'd have to model the knowledge of all agents. (To be continued. DEMO solution to be added!)