INFOMLSAI Logics for Safe AI
Coursework 2

Coursework released:	9 May 2023, on Blackboard
Coursework due:	23:59 19 May 2023, on Blackboard
Submission format:	a folder containing a pdf and an .ispl file, one per group

Please do the coursework in groups. Submit a single zipped folder on Blackboard for your group. The folder should contain a pdf file and the ispl file(s). Please name the folder group-X, where X is the number of your group, and state in the pdf file and comments in the ispl file the names of the members of the group.

Tasks that can be done in Week 3 (w/c 8 May)

The following tasks can be done after the lectures on Epistemic Logic and Epistemic Logic with Common and Distributed Knowledge.

W3-1 Consider three agents, a, b and c who each were dealt one of the following cards: a card with one dot, a card with two dots, or a card with three dots. Each agent knows their card but not the cards of the other agents. This scenario was described in the lecture on epistemic logic, and an incomplete Kripke model corresponding to it is shown in the slides. Provide a completed Kripke model M_{abc} (specify the states, indistinguishability relation for all three agents, and a valuation). The propositions are $a1$ for agent a has the card with 1 dot, $a2$ for agent a has the card with 2 dots, etc. for $a3, b1, b2, b3, c1, c2, c3$.

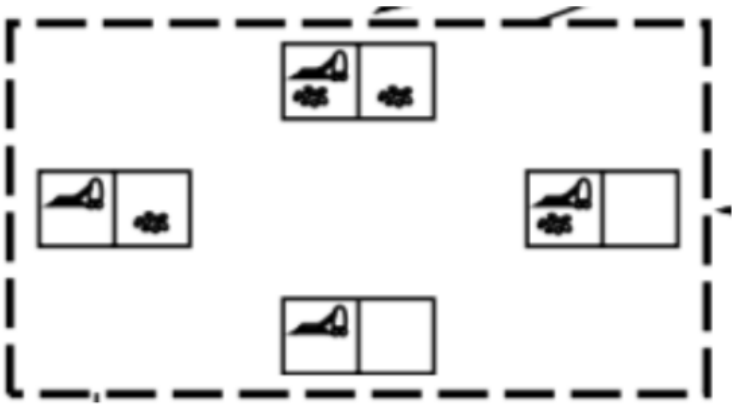
(1 point)

W3-2 In the model above, in the state q_{123} where a has the card with 1 dot, b has the card with 2 dots, and c has the card with 3 dots, does it hold that:

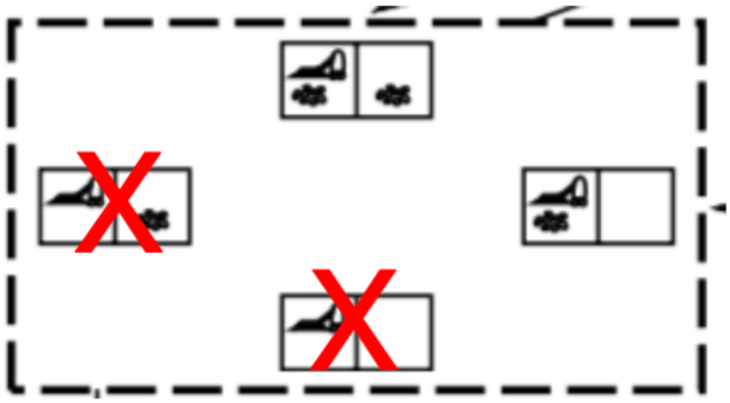
- it is distributed knowledge between a, b and c that $b2$. Express this is in $ELCD$ and argue why it is true (or false).
- it is common knowledge between a, b and c that $a1 \vee b2 \vee c3$. Express this is in $ELCD$ and argue why it is true (or false).
- it is it is common knowledge between a, b and c that $a1 \vee a2 \vee a3$. Express this is in $ELCD$ and argue why it is true (or false).

(1 point)

W3-3 A basic action in epistemic planning is *a truthful public announcement*: some formula φ is announced, and as a result, each agent in the system considers possible only the states where φ is true. For example, if the vacuum cleaner agent from Russell and Norwig's textbook that is aware of two rooms A and B (A on the left, B on the right) hears the truthful public announcement $\neg$cleanA, it will eliminate the states where cleanA holds, and will be left with just two possibilities:



⇓ announcement of $\neg$cleanA



Formally, a truthful public announcement of φ is applied to a pair of a Kripke model $M = (St, \{\sim_i : i \in Agt\}, \mathcal{V})$ and a state $q \in St$ such that $M, q \models \varphi$ and produces an updated Kripke model $M^\varphi = (St^\varphi, \{\sim_i^\varphi : i \in Agt\}, \mathcal{V}^\varphi)$ where $St^\varphi = St \cap \{q' \mid M, q' \models \varphi\}$, and for each i, $\sim_i^\varphi$ is $\sim_i$ restricted to the remaining states, and $\mathcal{V}^\varphi$ is $\mathcal{V}$ restricted to the remaining states. For more background, see the entry on Public Announcement Logic in the Stanford Encyclopaedia of Philosophy.

Question: In the Kripke model M_{abc} above, and the state q_{123}, Give a truthful public announcement φ such that after this announcement c knows which card each agent has, but a does not know which card c has?

Formally, is there a formula φ such that:

- $M_{abc}, q_{123} \models \varphi$
- $M^{\varphi}_{abc}, q_{123} \models K_c(a1 \wedge b2 \wedge c3)$
- $M^{\varphi}_{abc}, q_{123} \models \neg K_a c3$

If yes, give the formula and describe M^{φ}_{abc}.

(1 point)

W3-4 Write a formula that says: it is common knowledge between a and b that c knows which card each agent has. Is this formula true after the announcement of the formula φ from W3-3?

(1 point)

W3-5 Write a formula which is true in M_{abc}, q_{123} but when announced does not change the knowledge of any agent.

(1 point)

Tasks that can be done during Week 4 (w/c 15 May)

The following tasks can be done after the lectures on interpreted systems, CTLK, and model checking CTLK.

W4-1 Describe M_{abc} in ISPL. Make the distribution of cards a property of the environment, and each of the agents a, b and c can observe the value of its own variable. Make the initial states the set of all possible states of the system (card distributions). MCMAS related hints: recall that the Environment section of an ispl file can have an empty set of actions, empty protocol and empty evolution. Agents should have at least one local variable (could be dummy), one action (could be nil), a non-empty protocol, and some kind of evolution (of the dummy variable). (2 points)

W4-2 The following properties should be true in your model in all initial states. Translate them to MCMAS notation and add them to your file to check.

Everywhere globally, if $c3$ is true, then agent c knows $c3$

Everywhere globally, agent a does not know $c3$

Everywhere globally, it is not common knowlege between a and b that $c3$

Everywhere globally, it is distributed knowlege between a and b that $c3$

Everywhere globally, c knows that b does not know that $c3$

Everywhere globally, if a has card 1, then a knows that c has card 2 or 3.

Include in your submission the output of MCMAS when checking the properties above.

(1 point)

W4-4 Take your solution to modelling the submarine in MCMAS (or the model solution) and modify it so that the agent can observe whether the hatch is open or not, and whether it is sunk or not, but not whether it is on the surface or not. In the initial state, it should be true that the agent does not know that it is on the surface. Eventually the agent actually reaches a state when it does know that it is on the surface. Experiment with MCMAS couterexample generator to find out how it does come to know that it is on the surface. Write a brief explanation and include the MCMAS output.

(2 points)