INFOMLSAI Logics for Safe AI
Coursework 2 Model Answers

Coursework released:	10 May 2021, on Blackboard
Coursework due:	23:59 21 May 2021, on Blackboard
Submission format:	a folder containing a pdf and an .ispl file, one per group

Please do the coursework in groups of 2-3 people. Submit a single zipped folder on Blackboard for your group. The folder should contain the pdf file and the ispl file. 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 10 May)

The following tasks can be done after watching 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 $a3, b1, b2, b3, c1, c2, c3$.

Answer:

$M_{abc} = (\{q_{123}, q_{132}, q_{213}, q_{231}, q_{312}, q_{321}\}, \{\sim_i \mid i \in \{a, b, c\}\}, \mathcal{V})$ where

- naming convention for states: q_{ijk} means a has card i, b has card j and c has card k
- $\sim_i$ connects states where i has the same card, for example $q_{123} \sim_a q_{132}$, $q_{213} \sim_a q_{231}$, $q_{312} \sim_a q_{321}$, and $q_{123} \sim_b q_{321}$.
- $q_{ijk} \in \mathcal{V}(ai)$, $q_{ijk} \in \mathcal{V}(bj)$, $q_{ijk} \in \mathcal{V}(ck)$

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 $a1 \wedge b2 \wedge c3$. Express this is in $ELCD$ and argue why it is true (or false).

  Answer: $D_{\{1,2,3\}}(a1 \wedge b2 \wedge c3)$. This formula is true because the only state that is in $\sim^D_{\{1,2,3\}}$ (which is $\sim_1 \cap \sim_2 \cap \sim_3$) to q_{123} is q_{123} itself, and the formula $a1 \wedge b2 \wedge c3$ is true there.

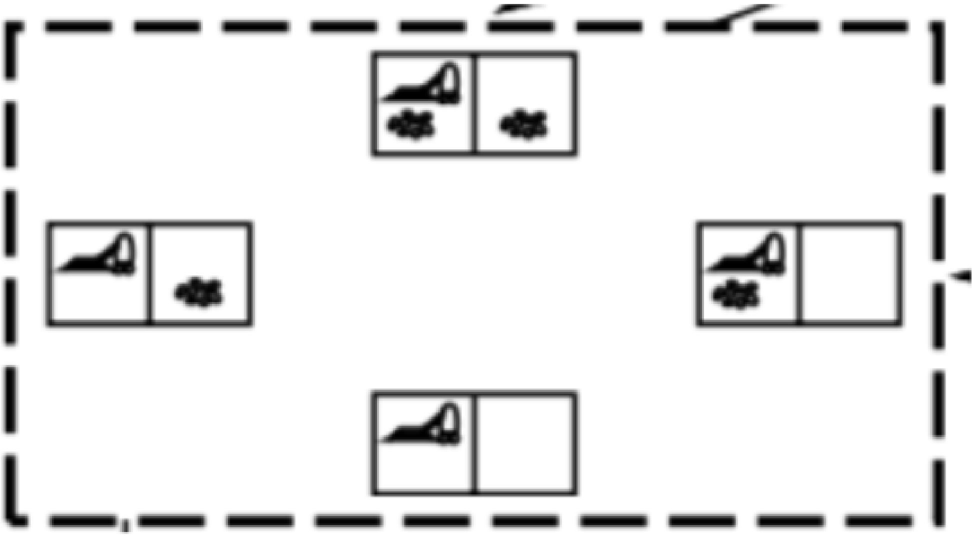
- it is common knowledge between a, b and c that $a1 \vee b2 \vee c3$. Express this is in $ELCD$ and argue why it is true (or false).

  Answer: $C_{\{a,b,c\}}(a1 \vee b2 \vee c3)$ is false in q_{123} because all states are reachable from q_{123} by $\sim^C_{\{a,b,c\}}$, that is, by a path of $\sim_a$, $\sim_b$ and $\sim_c$ relations, including states where $a1 \vee b2 \vee c3$ is false, for example q_{231}.

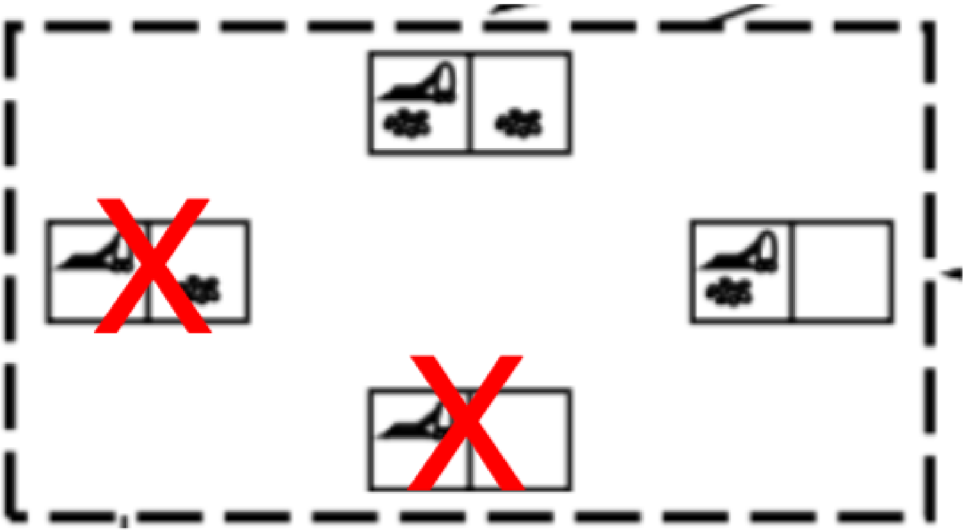
- it is common knowledge between a, b and c that $a1 \vee a2 \vee a3$. Express this is in $ELCD$ and argue why it is true (or false).

  Answer: $C_{\{a,b,c\}}(a1 \vee a2 \vee a3)$ is true in q_{123} because in all states $a1 \vee a2 \vee a3$ is true, so it is true in all states reachable by $\sim^C_{\{a,b,c\}}$.

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 hears the truthful public announcement ¬cleanA, it will eliminate the states where cleanA holds, and will be left with just two possibilities:



⇓ announcement of ¬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 $M, q \models \varphi$ [1] 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}, is there a truthful public announcement φ such that after this announcement b and c know which card each agent has but a does not know which cards b and c have?

Formally, is there a formula φ such that:

- $M_{abc}, q_{123} \models \varphi$
- $M_{abc}^\varphi, q_{123} \models E_{\{b,c\}}(a1 \wedge b2 \wedge c3)$
- $M_{abc}^\varphi, q_{123} \models \neg K_a(b2 \wedge c3)$

If yes, give the formula and describe M_{abc}^φ. Does φ also guarantee common knowledge $C_{\{b,c\}}(a1 \wedge b2 \wedge c3)$?

Answer: yes, for example a formula $\neg b1 \wedge \neg c1$ ($a1$ is even better):

- $M_{abc}, q_{123} \models \neg b1 \wedge \neg c1$
- $M_{abc}^{\neg b1 \wedge \neg c1}, q_{123} \models E_{\{b,c\}}(a1 \wedge b2 \wedge c3)$
- $M_{abc}^{\neg b1 \wedge \neg c1}, q_{123} \models \neg K_a(b2 \wedge c3)$

[1] (M, q) is called a pointed Kripke model

The first buller point is obvious. The second is because the announcement eliminates uncertainty for b and c about whether the other agent has card 1.In $M^{\neg b1 \wedge \neg c1}_{abc}$, the only state that is indistinguishable by b or c from q_{123} is q_{123} itself so they both know $a1 \wedge b2 \wedge c3$. The third bullet point is true because the announcement does not eliminate any uncertainty for a, because a has card 1 and does not consider states possible where b has 1 or c has 1. So it still considers possible a state where $b3 \wedge c2$ is true and $b2 \wedge c3$ is false.

Tasks that can be done during Week 4 (w/c 17 May)

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

W4-1 Represent the scenario of two robots and a carriage where the first robot can detect the colour of the surface and the second robot the texture, as an interpreted system in ISPL. The scenario is like the one in the reader on p.44, but both robots can push the carriage. Robot 1 pushes clockwise, and robot 2 pushes counterclockwise. If both robots push at the same time, the carriage does not move. Define group `g` as $\{$`Robot1, Robot2`$\}$. The following formulas should be true in the initial state q_0:

Formula	**ISPL Property**
$\mathsf{pos2} \rightarrow \neg K_1 \mathsf{pos2}$	`pos2 -> !K(Robot1,pos2);`
$\mathsf{pos2} \rightarrow K_1 \neg \mathsf{pos1}$	`pos2 -> K(Robot1,!pos1);`
$\mathsf{pos2} \rightarrow K_2 K_1 \neg \mathsf{pos1}$	`pos2 -> K(Robot2,K(Robot1,!pos1));`
$\mathsf{pos2} \rightarrow D_{\{1,2\}} \mathsf{pos2}$	`pos2 -> DK(g,pos2);`
$\mathsf{pos2} \rightarrow \neg E_{\{1,2\}} \mathsf{pos2}$	`pos2 -> !GK(g,pos2);`
$\mathsf{pos2} \rightarrow E_{\{1,2\}} \neg \mathsf{pos1}$	`pos2 -> GK(g,!pos1);`
$\mathsf{pos2} \rightarrow \neg C_{\{1,2\}} \mathsf{pos2}$	`pos2 -> !GCK(g,pos2);`
$\mathsf{pos2} \rightarrow \neg C_{\{1,2\}} \neg \mathsf{pos1}$	`pos2 -> !GCK(g,!pos1);`
$AGC_{\{1,2\}}(\mathsf{pos0} \vee \mathsf{pos1} \vee \mathsf{pos2})$	`AG GCK(g,(pos0 or pos1 or pos2))`
$EFE_{\{1,2\}} \neg \mathsf{pos1}$	`EF GK(g,!pos1)`

Answer: ISPL file attached.

W4-2 Give a model checking algorithm for CTLK that uses only the existential pre-image function (i.e., universal properties are checked without using the universal pre-image).

Answer: The simplest way to achieve this is to rewrite all formulas that require universal pre-image in dual or existential form. The lectures showed how to do this for epistemic modalities. Temporal modalities can be equivalently rewritten as follows:

$AX\varphi$	as	$\neg EX \neg \varphi$
$AG\,\varphi$	as	$E\top\, U \neg \varphi$
$A\varphi\, U\, \psi$	as	$\neg EG\, \neg \psi \wedge \neg E \neg \psi\, U\, (\neg \varphi \wedge \neg \psi)$

The first two duals are fairly straightforward. The third dual is a correct translation because: the first conjunct $\neg EG\, \neg \psi$ says that on all paths ψ does eventually

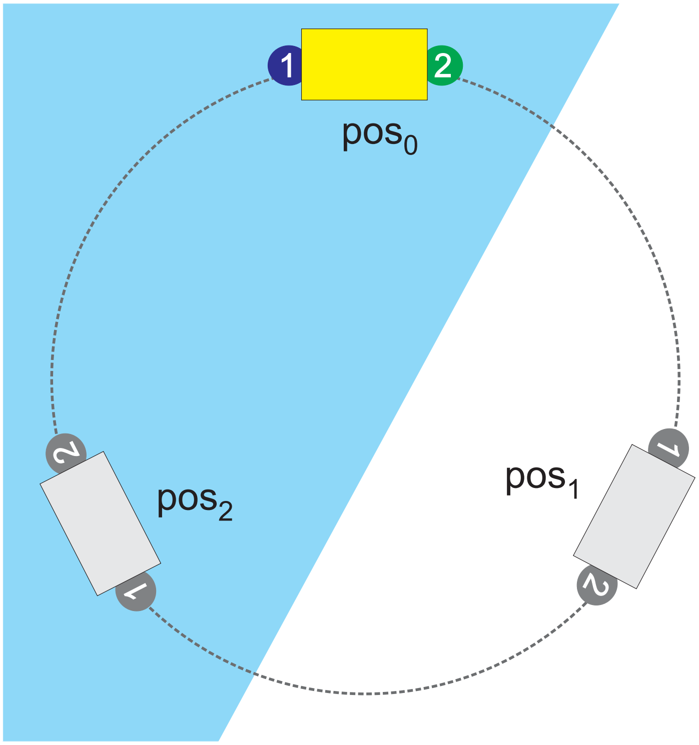


Figure 1: Two robots and a carriage

happen; the second conjunct says that there is no path where $\neg\psi$ holds until $\neg\varphi \wedge \neg\psi$ holds, which is the same as saying that on all paths either ψ happens straight away, or φ holds until it happens.

Another approach is to redefine the universal pre-image $pre_{\forall}(\rightarrow, Q)$ in terms of the existential pre-image as $St \setminus pre_{\exists}(\rightarrow, St \setminus Q)$, and substitute this computation in the algorithm. The definition is correct because $pre_{\exists}(\rightarrow, St \setminus Q)$ returns all states that have at least one $\rightarrow$ successor in the complement of Q. The complement of this set is the set of states all of whose successors are in Q, which is exactly $pre_{\forall}(\rightarrow, Q)$.