# INFOMLSAI Logics for Safe AI
# Final Exam

**This is an open-book examination**

*You can use your notes, the textbook, lecture slides and recordings etc.*

## Answer ALL questions

*Marks available for each question part are shown in brackets in the right-hand margin. This exam is marked out of 100.*

*The suggested time to answer all four questions is about 2 hours.*
*You can write your answers using word processing software, and also include within the document scanned or photographed portions that you have written by hand, or write it all by hand and scan or photograph it to produce a single PDF.*

*Use the standard naming convention for your submission:* `YourSolisId.pdf`.
*Write your Solis ID number at the top of each page of your answers. Do not include your name.*

**This is an individual assessment. Under no circumstances are you to discuss any aspect of this assessment with anyone; nor are you allowed to share this document, ideas or solutions with others using email, social media, instant messaging, websites, or any other means. Your attempts at these questions must be entirely your own work. Those found to have collaborated with others will receive a mark of 0.**

# Supplementary Material

### The language of LTL:

$$\varphi := \top \mid p \mid \neg\varphi \mid \varphi \wedge \psi \mid \varphi \vee \psi \mid \varphi \rightarrow \psi \mid X\varphi \mid F\varphi \mid \varphi U\, \psi \mid G\, \varphi$$

where $\top$ is an always true formula and $p$ is a propositional variable.

### The language of CTL:

$$\varphi := \top \mid p \mid \neg\varphi \mid \varphi \wedge \psi \mid \varphi \vee \psi \mid \varphi \rightarrow \psi \mid EX\varphi \mid EF\varphi \mid E\varphi U\, \psi \mid EG\, \varphi \mid AX\varphi \mid AF\varphi \mid A\varphi U\, \psi \mid AG\, \varphi$$

where $\top$ is an always true formula and $p$ is a propositional variable.

### The language of ELCD:

$$\varphi := \top \mid p \mid \neg\varphi \mid \varphi \wedge \psi \mid \varphi \vee \psi \mid \varphi \rightarrow \psi \mid K_i\varphi \mid E_A\varphi \mid C_A\varphi \mid D_A\varphi$$

where $\top$ is an always true formula, $p$ is a propositional variable, and $A$ is a set of agents.

### The language of ATL:

$$\varphi := \top \mid p \mid \neg\varphi \mid \varphi \wedge \psi \mid \varphi \vee \psi \mid \varphi \rightarrow \psi \mid \langle\langle A\rangle\rangle X\varphi \mid \langle\langle A\rangle\rangle F\varphi \mid \langle\langle A\rangle\rangle \varphi U\, \psi \mid \langle\langle A\rangle\rangle G\, \varphi$$

where $\top$ is an always true formula, $p$ is a propositional variable, and $A$ is a set of agents.

**Q1** Consider the state transition system $M$ in Figure 1. Propositions are false in states unless indicated to be true, e.g., $p$ is false in state $s_1$ and true in $s_2$.

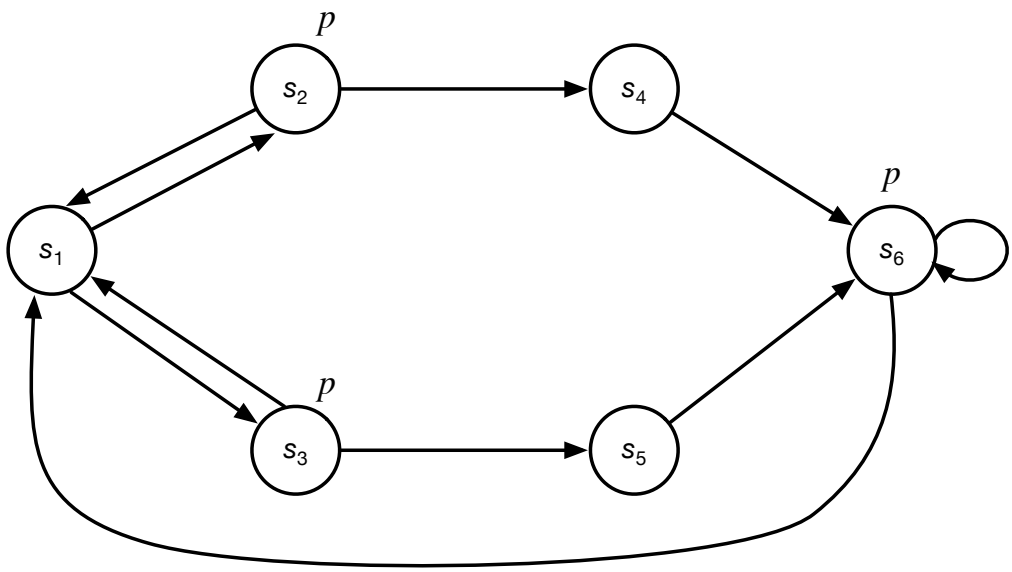


Figure 1: State transition system $M$.

(i) Express in LTL: $\neg p$ holds now, and it holds globally that if $p$ is true, then in the next state $\neg p$ is true. Is this formula true on all paths starting in $s_1$? Explain why with reference to the truth definition for LTL. (5 marks)

(ii) Express in CTL: it is not possible to reach a state where in all states on all paths in the future $p$ holds. Is this formula true in $s_1$? Explain why with reference to the truth definition for CTL. (5 marks)

(iii) What does the formula $AG\,(\neg p \vee EXp)$ mean? Is it true in $s_1$? Explain why with reference to the truth definition for CTL. (5 marks)

(iv) What does the formula $A(p \vee \neg p)\,U\,EG\,p$ mean? Is it true in $s_1$? Explain why with reference to the truth definition for CTL. (5 marks)

(v) Trace the CTL global model checking algorithm for the formula $E\top\,U\,AXp$, where $\top$ is a tautology (formula which is true everywhere) on the state transition system $M$. That is, list the values of the variables at each step of the algorithm until the algorithm terminates (see the model answer to the mock exam). Use the algorithm presented in Lecture 2/2 (slides 7–10). (5 marks)

**Q2** Consider the Kripke model $M_{kripke}$ in Figure 2. Possible worlds or states are nodes labelled by two propositional variables, for example $\mathsf{fish_a}$ and $\mathsf{fish_b}$. ($\mathsf{X}_a$ means that $a$ is having $\mathsf{X}$ for dinner, and $\mathsf{Y}_b$ means that $b$ is having $\mathsf{Y}$ for dinner.)

A line with a label $a$ denotes an indistinguishability relation $\sim_a$. Note that not all relations are depicted, since $\sim_a$ is also reflexive (each state is connected to itself by $\sim_a$) and transitive (if $\mathsf{fish_a fish_b}$ is connected to $\mathsf{fish_a meat_b}$, and $\mathsf{fish_a meat_b}$ is connected to $\mathsf{fish_a veg_b}$, then $\mathsf{fish_a fish_b}$ is connected by $\sim_a$ to $\mathsf{fish_a veg_b}$). Same for $\sim_b$.

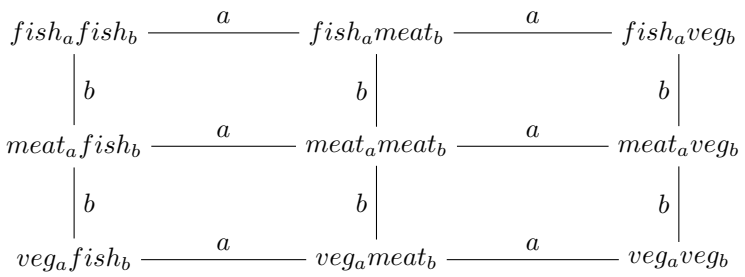


Figure 2: Kripke model $M_{kripke}$

(i) Represent this model with states (possible worlds) $w_1, \ldots, w_9$, two agents $a$ and $b$ with indistinguishability relations $\sim_a$ and $\sim_b$, and propositions $\mathsf{fish_a}$, $\mathsf{fish_b}$, $\mathsf{meat_a}$, $\mathsf{meat_b}$, $\mathsf{veg_a}$, $\mathsf{veg_b}$, which are true in those states where they are part of the label in Figure 2. States are numbered from the top row and left to right, so that the state in top left corner is $w_1$ labelled with $\mathsf{fish_a fish_b}$. (2 marks)

(ii) What are the states reachable by $\sim^D_{a,b}$ (distrbuted knowledge) relation from $w_1$ ($\mathsf{fish_a fish_b}$)? What are the states reachable by $\sim^C_{a,b}$ (common knowledge) relation from $w_1$ ($\mathsf{fish_a fish_b}$)? (5 marks)

(iii) Express in epistemic logic: agent $a$ knows that he does not know what agent $b$ is having for dinner (fish, meat, or veg). Is this formula true in $w_1$? Justify your answer with reference to the truth definition for epistemic logic formulas in Kripke models.

(5 marks)

(iv) Express in epistemic logic: it is common knowledge between agents $a$ and $b$ that agent $a$ knows what he is having for dinner. Is this true in $w_1$? Justify your answer with reference to the truth definition for epistemic logic formulas in Kripke models. (5 marks)

(v) Model the following scenario as a Kripke model. There are three agents 1, 2 and 3, and each of them is either busy or not. Agent 1 knows whether it is busy, and can also see agents 2 and 3 and knows whether they are busy or not. Agent 2 knows whether it is busy, and can see agent 3 and knows whether agent 3 is busy (but does not know about agent 1). Agent 3 only knows about its own business but not about agents 1 and 2. Assume propositions $\mathsf{busy_1}$, $\mathsf{busy_2}$, $\mathsf{busy_3}$. (8 marks)

**Q3** Consider the concurrent game structure $M_3$ in Figure 3. There are two agents, $a$ and $b$. Each agent has two actions in each state, 0 and 1. Transitions are by a pair of actions by $a$ and $b$, as indicated in edge labels, for example from $q_0$ there is a transition by $(0,0)$ to $q_2$ and by $(0,1)$ to $q_1$. There is a propositional variable $p$ that is true only in $q_1$.

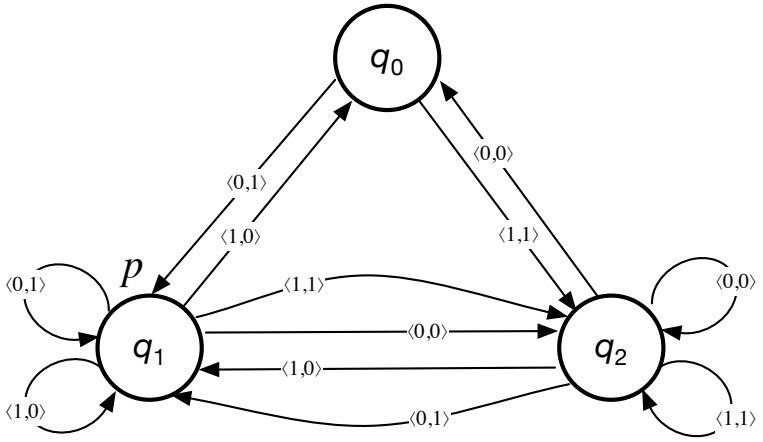


Figure 3: CGS $M_3$.

(i) Specify each component in $M_3 = \langle Agt, St, \mathcal{V}, Act, d, o\rangle$ based on Figure 3. (2 marks)

(ii) Express in ATL: agent $a$ has a strategy to make $p$ true at some point in the future. Is this formula true in $q_0$? Justify your answer with reference to the truth definition for ATL. If the formula is true, give a witness strategy.) (5 marks)

(iii) Express in ATL: agents $a$ and $b$ have a strategy to make $p$ false forever. Is this formula true in $q_0$? Justify your answer with reference to the ATL semantics. If the formula is true, give a witness strategy. (5 marks)

(iv) Suppose a modality $B$ with the following truth definition is added to ATL:

$M, q \models \langle\langle A\rangle\rangle B\varphi$ if, and only if, there is a collective strategy $s_A$ such that, for every path $\lambda \in out(q, s_A)$, there exists an $i \geq 0$, such that for all $j \geq i$, $M, \lambda[j] \models \varphi$.

Give pseudocode for the case in the global model checking algorithm for $\langle\langle A\rangle\rangle B\varphi$; use the algorithm presented in Lecture 7/1 (slides 4–5) as a template. (7 marks)

(v) Consider the algorithm ALG below that takes as input a CGS $M = \langle Agt, St, \mathcal{V}, Act, d, o\rangle$, a state $q \in St$, a set of states $Q \subseteq St$ and a non-empty coalition $A \subseteq Agt$, and returns true or false. Below, $d_A(q)$ is the set of joint actions by $A$ from $q$ and $d_{Agt}(q)$ is the set of joint actions by all agents from $q$. For a joint action by all agents $\alpha \in d_{Agt}(q)$, $\alpha_A$ is the sub-action of $\alpha$ which consists of actions by $A$. Assume that for actions $\sigma \in d_A(q)$ and $\alpha \in d_{Agt}(q)$ it is possible to check in constant time whether $\sigma = \alpha_A$. It is also possible to check in constant time whether $o(q, \alpha) \in Q$ holds or not.

```
function ALG(M, q, Q, A)
    for σ ∈ d_A(q) do
        check ← true
        for α ∈ d_Agt(q) do
            if σ = α_A and o(q, α) ∉ Q then
                check ← false
        if check then
            return true
    return false
```

State the big-O complexity of this algorithm in terms of the number of transitions $m$ in $M$ (that is, the size of the table defining $o$). Is it polynomial or exponential in $m$? (6 marks)

**Q4** Consider the concurrent epistemic game structure $M_4$ in Figure 4. There are two agents, $a$ and $b$. Each agent has two actions, 0 and 1, which are both available in $s_1$ and $s_2$; in $s_3$ and $s_4$ agent $a$ also has both actions but agent $b$ has only action 0, and in $s_5$ and $s_6$ both agents only have action 0. Transitions are by a pair of actions by $a$ and $b$, as indicated in edge labels, for example from $s_1$ there is a transition by $(0, 1)$ to $s_3$ and by $(1, 1)$ to $s_4$. Indistinguishability relations are indicated by a dotted line with the agent's name, for example $s_5$ and $s_6$ are indistinguishable for $b$. There is a propositional variable $p$ that is true only in $s_5$.

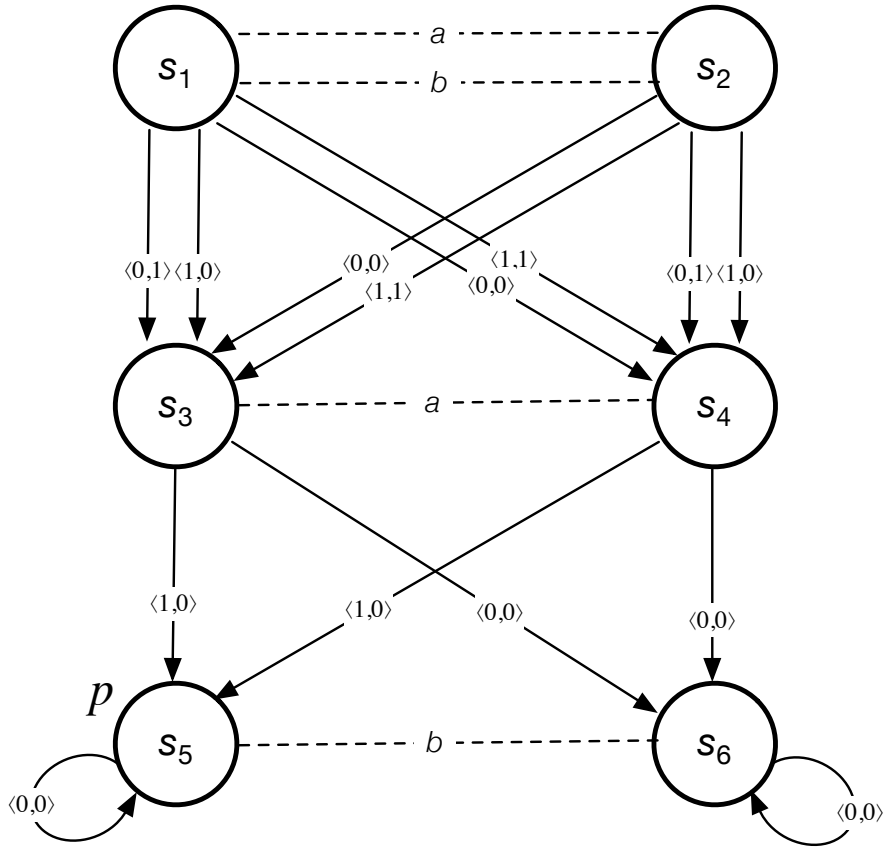


Figure 4: CEGS $M_4$.

(i) Specify each component in $M_4 = \langle Agt, St, \{\sim_i |\ i \in Agt\}, \mathcal{V}, Act, d, o\rangle$ based on Figure **??**. (2 marks)

(ii) Is the formula $\langle\langle a\rangle\rangle Fp$ true in $M_4, s_1$ under ATL$_{ir}$ semantics? Justify your answer with reference to the truth definition for ATL$_{ir}$. If the formula is true, give a witness strategy. (5 marks)

(iii) Is the formula $\langle\langle a, b\rangle\rangle F\langle\langle b\rangle\rangle Xp$ true in $s_1$ under ATL$_{ir}$ semantics? Justify your answer with reference to the truth definition for ATL$_{ir}$. If the formula is true, give a witness strategy. (5 marks)

(iv) In an interpreted system corresponding to $M_4$, how many local states for agents $a$ and $b$ would there be and why? (5 marks)

(v) Consider the problem of checking whether $M, q \models_{ir} \langle\langle A\rangle\rangle X\varphi$, where $M$ is A CEGS and $q$ a state in $M$. Explain why finding a joint action by $A$ in $q$ and checking whether the same action enforces $\varphi$ from all $q' \sim_A^E q$ does not correspond to the truth definition for ATL$_{ir}$ semantics if $A$ contains more than one agent. (8 marks)