Logics for Safe AI | Exam

6324363, Utrecht University, The Netherlands

1 Q1

1.1 Express in LTL

$$\neg pUpX\neg p \tag{1}$$

This formula holds on all paths starting in s_1 because iff $\lambda[1..\infty] \models \varphi$.

1.2 Express in CTL

$$\neg E(AGp) \tag{2}$$

This formula is not true in s_1 because iff for all paths λ, starting from q, we have $M, \lambda \models \varphi$.

1.3 CTL to English

On all paths, the following is true in all future moments - **p** is not true and there exists a path to a state where **p** is true. This formula is not true in s_1 because iff for all paths λ, starting from q, we have $M, \lambda \models \varphi$. When looking at system, there are future states where p holds.

1.4 CTL to English

On all paths, **p** is true and **p** is not true until there exists a path which for all future moments leads to a state where **p** is true. This formula is not true in s_1 because iff for all paths λ, starting from q, we have $M, \lambda \models \varphi$. When looking at the system, there are no states with unspecified p.

1.5 Tracing formulas

Let us call the transition system M, with the valuation $V(p) = s2, s3, s6$ as follows

$$[p]_M \leftarrow s_2, s_3, s_6 \tag{3}$$

Computing $[AXp]_M$:

$$[AXp]_M \leftarrow pre_\forall(\{s_2, s_3, s_6\}) = \{s_1, s_4, s_5\} \tag{4}$$

Computing $[E\top UAXp]_M$:

$$\begin{aligned} &Q_1 \leftarrow \emptyset; Q_2 \leftarrow \{s_2, s_3, s_6\} \\ &Q_1 \leftarrow \{s_2, s_3, s_6\}, Q_2 \leftarrow pre_\exists(Q1) \cap \top \\ &[E\top UAXp]_M \leftarrow Q_1 = \{s_2, s_3, s_6\} \end{aligned} \tag{5}$$

2 Q2

2.1 Describing the Kripke model

First, to define model M_{kripke}, all the possible states need to be described. Let the states be St_{kripke} such that $St_{kripke} = \{w_0, w_1, ..., w_9\}$, where

- $w_1 = \{fish_a fish_b\}$;
- $w_2 = \{fish_a meat_b\}$;
- $w_3 = \{fish_a veg_b\}$;
- $w_4 = \{meat_a fish_b\}$;
- $w_5 = \{meat_a meat_b\}$;
- $w_6 = \{meat_a veg_b\}$;
- $w_7 = \{veg_a fish_b\}$;
- $w_8 = \{veg_a meat_b\}$;
- $w_9 = \{veg_a veg_b\}$.

2.1.1 *Indistinguishability relations.*

The states with indistinguishable knowledge for each agent $A = \{a, b\}$ have been described

- $a = \{w_1, w_2, w_3\}, \{w_4, w_5, w_6\}, \{w_7, w_8, w_9\}$;
- $b = \{w_1, w_4, w_7\}, \{w_2, w_5, w_8\}, \{w_3, w_6, w_9\}$.

2.1.2 *Valuation.*

Following the example given in course, the valuations can be described as follows. Let the valuations be the form of $x_y \mapsto \{St_{kripke}\}$, where

- $fish_a fish_b \mapsto \{w_1\}$;
- $fish_a meat_b \mapsto \{w_2\}$;
- $fish_a veg_b \mapsto \{w_3\}$;
- $meat_a fish_b \mapsto \{w_4\}$;
- $meat_a meat_b \mapsto \{w_5\}$;
- $meat_a veg_b \mapsto \{w_6\}$;
- $veg_a fish_b \mapsto \{w_7\}$;
- $veg_a meat_b \mapsto \{w_8\}$;
- $veg_a veg_b \mapsto \{w_9\}$.

2.2 Reachable states

2.3 Express in Epistemic logic

Agent a knows that he does not know what agent b is having for dinner (fish,meat, or veg):

$$K_a(\neg K_a fish_b \vee \neg K_a meat_b \vee \neg K_a veg_b) \tag{6}$$

This formula is true in w_1 as in all states, the states with different dinner options for agent b are indistinguishable for agent a.

More formally, a Kripke model $M = \langle St, \sim_i (i \in Agt), V\rangle$ consists of a non-empty set of states St, a valuation of propositions $V : PV \to 2^{St}$ and an indistinguishability relation $\sim_i$ for each agent i which informs the answer above.

2.4 Express in Epistemic logic

It is common knowledge between agents a and b that agent a knows what he is having for dinner.

$$C_{a,b}(K_a fish_a \vee K_a meat_a \vee K_a veg_a) \tag{7}$$

True, because $\sim^E_{a,b} = \bigcup_{i \in \{a,b\}} \sim_i$. For this relation, every state is in the same equivalence class. The transitive closure and therefore $\sim^C_{a,b}$ contains the same relations (all states "connected"). In all states $K_a fish_a \vee K_a meat_a \vee K_a veg_a$ is true. Therefore $C_{a,b}(K_a fish_a \vee K_a meat_a \vee K_a veg_a)$ is true in the model.

Author's address: 6324363, , Utrecht University, P.O. Box 80125, Utrecht, Utrecht, The Netherlands, 3508 TC.

2.5 Describing the Kripke model

First, to define model M_{kripke}, all the possible states need to be described. Let the states be St_{kripke} such that $St_{kripke} = \{w_0, w_1, ..., w_7\}$, where

- $w_1 = \{busy_1\}$;
- $w_2 = \{\neg busy_1\}$;
- $w_3 = \{busy_2\}$;
- $w_4 = \{\neg busy_2\}$;
- $w_5 = \{busy_3\}$;
- $w_6 = \{\neg busy_3\}$;
- $w_7 = \{busy_1 busy_2\}$;
- $w_8 = \{\neg busy_1 busy_2\}$;
- $w_9 = \{busy_1 \neg busy_2\}$;
- $w_{10} = \{\neg busy_1 \neg busy_2\}$;
- $w_{11} = \{busy_1 busy_3\}$;
- $w_{12} = \{\neg busy_1 busy_3\}$;
- $w_{13} = \{busy_1 \neg busy_3\}$;
- $w_{14} = \{\neg busy_1 \neg busy_3\}$;
- $w_{15} = \{busy_2 busy_3\}$;
- $w_{16} = \{\neg busy_2 busy_3\}$;
- $w_{17} = \{busy_2 \neg busy_3\}$;
- $w_{18} = \{\neg busy_2 \neg busy_3\}$;
- $w_{19} = \{busy_1 busy_2 busy_3\}$;
- $w_{20} = \{\neg busy_1 busy_2 busy_3\}$;
- $w_{21} = \{busy_1 \neg busy_2 busy_3\}$;
- $w_{22} = \{busy_1 busy_2 \neg busy_3\}$;
- $w_{23} = \{\neg busy_1 \neg busy_2 busy_3\}$;
- $w_{24} = \{busy_1 \neg busy_2 \neg busy_3\}$;
- $w_{25} = \{\neg busy_1 busy_2 \neg busy_3\}$;
- $w_{26} = \{\neg busy_1 \neg busy_2 \neg busy_3\}$;

2.5.1 *Indistinguishability relations.*

The states with indistinguishable knowledge for each agent $A = \{1, 2\}$ have been described

- $1 = \{w_1\}, \{w_2\}, \{w_3\}, \{w_4\}, \{w_5\}, \{w_6\}, \{w_7\}.\{w_8\}, \{w_9\}, \{w_{10}\}, \{w_{11}\}, \{w_{12}\}, \{w_{13}\}, \{w_{14}\}, \{w_{15}\}, \{w_{16}\}, \{w_{17}\}, \{w_{18}\}, \{w_{19}\}, \{w_{20}\}, \{w_{21}\}, \{w_{22}\}, \{w_{23}\}, \{w_{24}\}, \{w_{25}\}, \{w_{26}\}$;
- $2 = \{w_3\}, \{w_4\}, \{w_5\}, \{w_6\}, \{w_{15}\}, \{w_{16}\}, \{w_{17}\}, \{w_{18}\}$;
- $3 = \{w_5\}, \{w_6\}$;

2.5.2 *Valuation.*

Following the example given in course, the valuations can be described as follows. Let the valuations be the form of $x_y \mapsto \{St_{kripke}\}$, where

- $busy_1 \mapsto \{w_1, w_7, w_9, w_{11}, w_{13}, w_{19}, w_{21}, w_{22}, w_{24}\}$;
- $\neg busy_1 \mapsto \{w_2, w_8, w_{10}, w_{12}, w_{14}, w_{20}, w_{23}, w_{25}, w_{25}\}$;
- $busy_2 \mapsto \{w_3, w_7, w_8, w_{15}, w_{17}, w_{19}, w_{20}, w_{22}, w_{25}\}$;
- $\neg busy_2 \mapsto \{w_4, w_9, w_{10}, w_{16}, w_{18}, w_{21}, w_{24}, w_{26}\}$;
- $busy_3 \mapsto \{w_5, w_{11}, w_{12}, w_{15}, w_{16}, w_{19}, w_{20}, w_{21}, w_{23}\}$;
- $\neg busy_3 \mapsto \{w_6, w_{13}, w_{14}, w_{17}, w_{18}, w_{22}, w_{24}, w_{25}, w_{26}\}$.

3 Q3

3.1 Describing the Concurrent Game Structure

Below, a description for concurrent game structure (CGS) is given. CGS incorporates multiple elements, include the set of agents and states and actions taken simultaneously, a valuation of propositions, specific actions available to a specific agent in a specific state and also a deterministic transition function that assigns outcome states to states and tuples of actions.

A concurrent game structure (CGS) is a tuple

$$M_{cgs} = (\{a, b\}, \{q_0, q_1, q_2\}, v, \{0, 1\}, d, o) \tag{8}$$

, where

V is defined as:

- $V(p) = \{q_1\}$

d is defined as:

- $d_{Agt}(q) = \{0, 1\}$;
- $\forall Agt \in \{a, b\}, q \in \{q_0, q_1, q_2\}$.

o is defined as:

- $o(q_0, 0, 0) = o(q_0, 1, 0) = -$;
- $o(q_0, 0, 1) = (q_1, 0, 1) = o(q_2, 0, 1) = o(q_2, 1, 0) = q_1$;
- $o(q_0, 1, 1) = o(q_1, 0, 0) = o(q_1, 1, 1) = o(q_2, 1, 1) = q_2$;
- $o(q_1, 1, 0) = \{q_0, q_1\}$;
- $o(q_2, 0, 0) = \{q_0, q_2\}$.

3.2 Express in ATL

Agent a has a strategy to make p true at some point in the future:

$$\langle\langle a \rangle\rangle Fp \tag{9}$$

This is untrue in q_0 because there is no such strategy that agent a can enforce on its own to satisfy the requirement of reaching state q_1.

$$M, q_0 \models \langle\langle a \rangle\rangle Fp \tag{10}$$

3.3 Express in ATL

Agents a and b have a strategy to make p false forever.

$$\langle\langle a, b \rangle\rangle G \neg p \tag{11}$$

This is true in q_0 because there is a strategy () that the coalition of agents a and b can enforce to satisfy the requirement of never reaching state q_1.

$$M, q_0 \models \langle\langle a, b \rangle\rangle G \neg p \tag{12}$$

Let s_1 be the strategy function for the coalition of agents a and b, where

- $s_1(q_0) = 1, 1$
- $s_1(q_2) = 0, 0, 1, 1$

The other states will never be reached if the coalitions of agents a and b plays this strategy. To achieve completeness, a definition of a witness strategy can be as follows:

- $s_1(q_i) = 1, 1$
- $s_1(q_2) = 0, 0$
- $\forall i \in \{0, 2\}$

3.4 Adding modalities and verifying them

3.5 Complexity

The algorithm $ALG(M, q, Q, A)$ has complexity $O(|St| * |\varphi|)$ and is executed for each generated strategy model M_i. Since there are maximum $|Act|^{St}$ different models, the total complexity is $O(|St| * |\varphi| * |Act|^{St})$.

4 Q4

Consider a CEGS M_4, where $(\mathbb{A}gt, St, \sim_i |i \in \mathbb{A}gt, V, Act, d, out)$ is a concurrent game structure, and $\sim_a$ are indistinguishability relations over St, one per agent a in $\mathbb{A}gt$. We can now define the CEGS as a tuple

$$M_4 = (\mathbb{A}gt, St, \sim_i |i \in \mathbb{A}gt, V, Act, d, out) \tag{13}$$

, where

$\mathbb{A}gt$ is defined as:

$$\mathbb{A}gt = \{a, b\} \tag{14}$$

St is defined as:

- $s_1 =$ a plays $\{0, 1\}$, b plays $\{0, 1\}$
- $s_2 =$ a plays $\{0, 1\}$, b plays $\{0, 1\}$
- $s_3 =$ a plays $\{0, 1\}$, b plays 0
- $s_4 =$ a plays $\{0, 1\}$, b plays 0
- $s_5 =$ a plays 0, b plays 0
- $s_6 =$ a plays 0, b plays 0

Indistinguishability relations are defined as:

- $\sim_a = \{s_1, s_2\}, \{s_3, s_4\}$
- $\sim_b = \{s_5, s_6\}$

V is defined as:

- $V(p) = \{s_5\}$

Actions are defined as:

- $Act = \{0, 1\}$

d is defined as:

- $d(\mathbb{A}gt, s_i) = \{0\}, \forall i \in \{1, ..., 6\}$
- $d(\mathbb{A}gt, s_i) = \{1\}, \forall i \in \{1, 2\}$
- $d(a, s_i) = \{1\}, \forall i \in \{3, 4\}$

o is defined as:

- $o(s_1, 0, 0) = o(s_1, 1, 1) = o(s_2, 0, 1) = o(s_2, 1, 0) = s_4$
- $o(s_1, 0, 1) = o(s_1, 1, 0) = o(s_2, 0, 0) = o(s_2, 1, 1) = s_3$
- $o(s_3, 0, 0) = o(s_4, 0, 0) = o(s_6, 0, 0) = s_6$
- $o(s_3, 1, 0) = o(s_4, 1, 0) = o(s_5, 0, 0) = s_5$
- $o(s_3, 0, 1) = o(s_3, 1, 1) = o(s_4, 0, 1) = o(s_4, 1, 1) = o(s_5, 0, 1) = o(s_5, 1, 0) = o(s_5, 1, 1) = o(s_6, 0, 1) = o(s_6, 1, 0) = o(s_6, 1, 1) = -$

4.1 Validation under ATL$_i r$

The formula $\langle\langle a\rangle\rangle Fp$ is untrue in M_4, s_1 under ATL$_i r$ sematics. This is because agent a can not enforce a uniform strategy from state s_1 to satisfy the requirement of reaching state s_5 for the proposition p to hold.

4.2 Validation under ATL$_i r$

4.3 Interpreted systems

In an interpreted system corresponding to M_4, there would be 4 local states for both agents a and b, namely

- a played 0, b played 0
- a played 0, b played 1
- a played 1, b played 0
- a played 1, b played 1

This is because each agents has its own individual view of the global state.

4.4 Explanation through truth definition

Under ATL$_i r$, there has to be a collective memoryless uniform strategy which works from all indistinguishable states. As strategy for A cannot be synthesized incrementally, we also can not specify which states are considered possible (strong uniformity).