INFOMLSAI Logics for Safe AI
Coursework 4 answers

Coursework released:	14 June 2021, on Blackboard
Coursework due:	23:59 22 June 2021, on Blackboard
Submission format:	a pdf file, one per group

CW4-1 Define a Concurrent Epistemic Game Structure (CEGS) $M_{chicken}$ for the following example. Two agents are moving towards each other. They are not sure whether they are in a country which drives on the left (q_0) or on the right (q_1). Each agent can execute actions $left$ and $right$. In state q_0, if both agents go left, they pass each other correctly. In q_1, if both agents go right, they pass each other correctly. For all other combinations of actions, there is a crash. List the states, agents, indistinguishability relations, actions, transition function, assignment. Use propositional variables lft for driving on the left and crash for there is a crash. Distinguish between states resulting from different combinations of actions and whether they are in a country that drives on the left or a country that drives on the right. The agents can observe whether a crash happened or not, and what the actions leading to it were. For example, if both agents went left and a crash happened, they both know that they were not in a left-driving country. (1 mark)

Answer:

$M_{chicken} = \langle \{1,2\}, \{q_0, q_1, q_2^l, q_2^r, q_3ll^r, q_3lr^l, q_3lr^r, q_3rl^l, q_3rl^r, q_3rr^l\}, \sim_1, \sim_2$ $, \mathcal{V}, \{left, right, nil\}, d, o\rangle$, where:

- q_0 and q_1 are the initial states, q_2^l and q_2^r are the left- and righthand side driving states where agents pass each other correctly. q_3ll^r is a state where a crash happened as a result of $\langle left, left\rangle$ in driving on the right country. q_3lr^l and q_3lr^r are the states where a crash happened as a result of $\langle left, right\rangle$ in left- and righ-driving countries, respectively. Similarly for q_3rl^l, q_3rl^r and q_3rr^l.
- $\sim_1$ and $\sim_2$ equivalence classes are $\{q_0, q_1\}$, $\{q_2^l\}$, $\{q_2^r\}$, $\{q_3ll^r\}$, $\{q_3lr^l, q_3lr^r\}$, $\{q_3rl^l, q_3rl^r\}$, $\{q_3rr^l\}$.
- $\mathcal{V}$ assigns sets of states to propositions:
  - $\mathcal{V}(\mathsf{crash}) = \{q_3ll^r, q_3lr^l, q_3lr^r, q_3rl^l, q_3rl^r, q_3rr\}$
  - $\mathcal{V}(\mathsf{lft}) = \{q_0, q_2l, q_3rr\}$

- $d(1, q_i) = d(2, qi_i) = \{left, right\}$ for $i \in \{0, 1\}$, $d(1, q_i) = d(2, q_i) = \{nil\}$ for $i \notin \{0, 1\}$
- o is as follows:
  - $o(q_0, \langle left, left\rangle) = q_2^l$;
  - $o(q_1, \langle right, right\rangle) = q_2^r$;
  - $o(q_0, \langle\alpha_1, \alpha_2\rangle) = q_3\alpha_1\alpha_2^l$ for $\langle\alpha_1, \alpha_2\rangle \neq \langle left, left\rangle$
  - $o(q_1, \langle\alpha_1, \alpha_2\rangle) = q_3\alpha_1\alpha_2^r$ for $\langle\alpha_1, \alpha_2\rangle \neq \langle right, right\rangle$
  - $o(q_i, nil) = q_i$ for $i \notin \{0, 1\}$.

CW4-2 Is it true in $M_{chicken}, q_0$ under the independent combination of ATL semantics with epistemic semantics (no uniform strategies requirement) that agent 1 has a memoryless strategy to enforce $\neg\mathsf{crash}$ in the next state ($\langle\langle 1\rangle\rangle X\neg\mathsf{crash}$)? (1 mark)

Answer: No, the formula $\langle\langle 1\rangle\rangle X\neg\mathsf{crash}$ is false in $M_{chicken}, q_0$. Under the independent combination of memoryless ATL semantics with epistemics,

$$M_{chicken}, q_0 \models \langle\langle 1\rangle\rangle X\neg\mathsf{crash}$$

iff (if and only if) there is a memoryless strategy s_1 for agent 1 such that for all paths $\lambda \in out(q_0, s_1)$ $M, \lambda[1] \models \neg\mathsf{crash}$. There are two possible actions that s_1 could assign to agent 1 in q_0. The first one is $left$, and the second one is $right$. If agent 1 chooses $left$ in q_0, agent 2 has two choices, and one of them (right) leads to a crash state. If agent 1 chooses $right$, a crash will result no matter what agent 2 does. So there is no strategy s_1 such that all paths generated by it satisfy $\neg\mathsf{crash}$ in the next state.

CW4-3 Does it hold under ATL$_{ir}$ semantics that $M_{chicken}, q_0 \models_{ir} \langle\langle 1\rangle\rangle X\neg\mathsf{crash}$? Explain your answer. (1 mark)

Answer: No. $M_{chicken}, q_0 \models_{ir} \langle\langle 1\rangle\rangle X\neg\mathsf{crash}$ iff there is a uniform memoryless strategy s_1 for agent 1 such that for all paths $\lambda \in \bigcup_{q' \sim_1 q_0} out(q', s_1)$,

$$M_{chicken}, \lambda[1] \models \neg\mathsf{crash}$$

So, since there is no strategy for agent 1 at all that enforces $\neg\mathsf{crash}$ in the next state (see CW4-2), then there can be no uniform strategy which enforces $\neg\mathsf{crash}$ in the next state from all $\sim_1$-indistinguishable states.

CW4-4 Is it true in $M_{chicken}, q_0$ under the independent combination of ATL semantics with epistemic semantics (no uniform strategies requirement) that both agents together have a memoryless strategy to enforce $\neg\mathsf{crash}$ in the next state ($\langle\langle 1, 2\rangle\rangle X\neg\mathsf{crash}$)? Explain your answer. (1 mark)

Answer:

Yes. $M_{chicken}, q_0 \models \langle\langle 1, 2\rangle\rangle X\neg\mathsf{crash}$ under the independent combination of memoryless ATL semantics with epistemic semantics iff there exists a joint strategy $s_{1,2}$ such that for all paths λ in $out(q_0, s_{1,2})$, $M_{chicken}, \lambda[1] \models \neg\mathsf{crash}$.

Such a strategy for 1 is $q_0 \mapsto left$, $q_1 \mapsto right$, and for the rest of the states q, $q \mapsto nil$, same for agent 2. This strategy is not uniform, because it assigns agent 1 different actions in $\sim_1$-indistinguishable states (and the same for agent 2), but uniformity is not required for this semantics. There is only one path from q_0 generated by this strategy, and the next state after $(left, left)$ on that path is q_2^l. Since $M_{chicken}, q_2^l \models \neg\mathsf{crash}$, we have $M_{chicken}, q_0 \models \langle\langle 1, 2\rangle\rangle X\neg\mathsf{crash}$.

CW4-5 Does it hold under ATL$_{ir}$ semantics that $M_{chicken}, q_0 \models_{ir} \langle\langle 1, 2\rangle\rangle X\neg\mathsf{crash}$? Explain your answer. (1 mark)

Answer:

No. $M_{chicken}, q_0 \models \langle\langle 1, 2\rangle\rangle X\neg\,\mathsf{crash}$ iff there exists a uniform memoryless strategy for 1 and 2, $s_{1,2}$, such that for all paths $\lambda \in \bigcup_{q' \sim^E_{1,2} q_0} \in out(q_0, s_{1,2})$, $M_{chicken}, \lambda[1] \models_{ir} \neg\mathsf{crash}$. There are 4 uniform strategies for agent 1 and agent 2 from q_0 (first strategy: 1 and 2 both choose $left$ in both q_0 and q_1; second strategy: 1 chooses $left$ in both q_0 and q_1 and 2 chooses $right$ in both q_0 and q_1; third strategy: 1 chooses $right$ in both q_0 and q_1 and 2 chooses $left$ in both q_0 and q_1; fourth strtegy: 1 and 2 both choose $right$ in both q_0 and q_1); in other states all strategies choose nil. The first strategy works from q_0, meaning that the only path generated by it satisfies $\neg\,\mathsf{crash}$ in the next state. But the same strategy should also work from all states $q' \sim^E_{1,2} q_0$, and $q_1 \sim^E_{1,2} q_0$. However the path generated by the first strategy from q_1 satisfies crash in the next state $q_3 ll^r$. The other strategies either don't work in both states or work from q_1 but not from q_0.

CW4-6 Is it true in $M_{chicken}, q_0$ under the independent combination of ATL semantics with epistemic semantics (no uniform strategies requirement) that both agents know that they have a memoryless strategy to enforce $\neg\mathsf{crash}$ in the next state ($K_1\langle\langle 1, 2\rangle\rangle X\neg\mathsf{crash} \wedge K_2\langle\langle 1, 2\rangle\rangle X\neg\mathsf{crash}$)? Explain your answer. (1 mark)

Answer:

Yes. In all states indistinguishable by $\sim_1$ from q_0 (q_0 itself and q_1) it holds that $\langle\langle 1, 2\rangle\rangle X\neg\mathsf{crash}$ (using a non-uniform strategy from CW4-4), and the same for all states indistinguishable from q_0 by $\sim_2$: in all q such that $q_0 \sim_2 q$, $M_{chicken}, q \models \langle\langle 1, 2\rangle\rangle X\neg\,\mathsf{crash}$. So it holds that $M_{chicken}, q_0 \models K_1\langle\langle 1, 2\rangle\rangle X\neg\,\mathsf{crash}$ and $M_{chicken}, q_0 \models K_2\langle\langle 1, 2\rangle\rangle X\neg\mathsf{crash}$, so $M_{chicken}, q_0 \models K_1\langle\langle 1, 2\rangle\rangle X\neg\mathsf{crash} \wedge K_2\langle\langle 1, 2\rangle\rangle X\neg\mathsf{crash}$.

CW4-7 Does it hold under ATL$_{ir}$ semantics that $M_{chicken}, q_0 \models_{ir} K_1\langle\langle 1, 2\rangle\rangle X\neg\mathsf{crash} \wedge K_2\langle\langle 1, 2\rangle\rangle X\neg\mathsf{crash}$? Explain your answer. (1 mark)

Answer:

No. From the answer to CW4-5 it follows that $M_{chicken}, q_0 \not\models_{ir} \langle\langle 1, 2\rangle\rangle X\neg\mathsf{crash}$. In fact in both states indistinguishable by 1 from q_0, $\langle\langle 1, 2\rangle\rangle X\neg\,\mathsf{crash}$ does not hold, because there is no uniform strategy such that the paths generated by this strategy from all indistnguishable states satisfy $X\neg\mathsf{crash}$. So $K_1\langle\langle 1, 2\rangle\rangle X\neg\mathsf{crash}$ does not hold, and similarly for $K_2\langle\langle 1, 2\rangle\rangle X\neg\mathsf{crash}$.

CW4-8 How would you say in ATL$_{ir}$ that agent 1 can ensure that eventually it knows whether it is in a left- or righhand side driving country? Is this formula true in q_0? Explain your answer. (1 mark)

Answer: $\langle\langle 1\rangle\rangle F(K_1\,\mathsf{lft} \vee K_1\neg\,\mathsf{lft})$. No, this is not true. Whichever action 1 chooses, it is possible that agent 2 performs a different action, and a crash results, but agent 1 will still not know whether he performed the correct action and the other agent the wrong action, or vice versa (agent 1 will not know whether he is in one of q_3lr^l, q_3lr^r if the actions were $(left, right)$, or it is in one of q_3rl^l, q_3rl^r if the actions were $(right, left)$).

CW4-9 How would you say in ATL$_{ir}$ that it is inevitable that if in the next state there is no crash, then agent 1 knows whether he is in a left- or righthand side driving country? Is this formula true in q_0? Explain your answer. (1 mark)

Answer: $\langle\langle\emptyset\rangle\rangle X(\neg\mathsf{crash} \to K_1\mathsf{lft} \vee K_1\neg\mathsf{lft})$. This formula is true because if the agent performs $left$ and there is no crash then lft must be true (the outcome is q_2^l), and similarly for performing $right$.

CW4-10 Give a model checking algorithm under ATL$_{ir}$ semantics for a language containing propositional variables, booleans, and formulas $\langle\langle a\rangle\rangle X^2\varphi$ where a is a single agent and $\langle\langle a\rangle\rangle X^2$ is a new modality which means 'reachable in two steps' (note that $\langle\langle a\rangle\rangle X^2$ is not definable in ATL$_{ir}$).

The truth definition for $\langle\langle a\rangle\rangle X^2\varphi$ is:

$M, q \models \langle\langle a\rangle\rangle X^2\varphi$ iff there is a memoryless uniform strategy s_a for a such that for all paths λ in $\bigcup_{q'\sim_a q} out(q', s_a)$, $M, \lambda[2] \models \varphi$.

(It requires that the strategy is guaranteed to enforce φ in two steps from any state indistinguishable from q.) What is the big O complexity of your algorithm as a function of the model size and formula size? (Note that we are not asking for the most efficient algorithm, just a correct one with correct complexity analysis.) (1 mark)

Answer: Suppose we are given M, q and $\langle\langle a\rangle\rangle X^2\varphi$. We are going to do local model checking (state by state). The simplest approach is to generate all memoryless strategies for a in M (all possible assignments of actions to states). Then remove from this set of strategies all non-uniform strategies that assign different actions in $\sim_a$-indistinguishable states. Each remaining uniform strategy can be made into a model M' by deleting all a's actions from M which do not conform to the strategy. In each M', all paths correspond to computations generated by a's uniform strategy. The model checking algorithm $mcheck(M', \varphi)$ used in each M' is the same as for CTL for the cases of propositional variables and boolean connectives. For the case $\varphi = \langle\langle a\rangle\rangle X^2\psi$, $mcheck(M', \varphi)$ returns $[\varphi]_{M'} = pre_\forall(pre_\forall([\psi]_{M'}))$. Finally, if $mcheck(M', \varphi)$ returns a set Q containing q, for strong uniformity, we need to check that all $q' \sim_a q$ are also in Q (that the strategy works from all states q' such that $q \sim_a q'$). If there is at least one M' where the set of states Q returned by $mcheck(M', \varphi)$ contains all states $\sim_a$-related to q, then we can return 'yes' on input M, q and $\langle\langle a\rangle\rangle X^2\varphi$.

Complexity: there are $O(d^n)$ memoryless strategies for a, where d is the number of $a's$ actions and n is the number of states (in the worst case, all d actions are possible in all states). To check each of strategy for uniformity, for each state q, we need to check that the same action is assigned to all states $q' \sim_a q$. Generating all uniform strategies takes $O(d^n \times | \sim_a | \times n)$ steps. To run the CTL-style model checking algorithm for each M' we need at most $O(|M'| \times |\varphi|)$ steps, which is dominated by $O(|M| \times |\varphi|)$ (since M' can be no larger than M). Checking for strong uniformity with respect to q in one M' requires $O(| \sim_a | \times n)$ steps (because we first have to find the neighbours of q in $\sim_a$, and then check if all neighbours are in Q, which in the worst case is of size n). The resulting complexity is $O(d^n \times | \sim_a |^2 \times n^2 \times |M| \times |\varphi|)$ or (removing dominated terms) $O(|M|^{|M|} \times |\varphi|)$, which is exponential in the size of the model.

References

[1] Alessio Lomuscio, Hongyang Qu, and Franco Raimondi. MCMAS: an open-source model checker for the verification of multi-agent systems. *Int. J. Softw. Tools Technol. Transf.*, 19(1):9–30, 2017.