# INFOMLSAI Logics for Safe AI
# Coursework 1

| | |
|---|---|
| **Coursework released:** | 26 April 2021, on Blackboard |
| **Coursework due:** | 23:59 7 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 folder on Blackboard for your group, indicating in the pdf file and in comments in the ispl file who the members of the group are.

## Tasks that can be done in Week 1 (w/c 26 April)

The first task is just to get started, and does not form part of the submission. Read the guide to installing MCMAS on Blackboard, and download and install MCMAS. Download the file `one-robot+carriage.ispl` from Blackboard (you can use this as a template for the tasks in Week 2). Check that everything works correctly by running MCMAS with `one-robot+carriage.ispl` as input. If you have problems, there will be a practical session on Wednesday 28th April where you can ask questions.

The following tasks can be done after watching the lectures on Temporal Logic and LTL, and form part of your coursework submission.

**W1-1** The *office world* domain is shown in Figure 1 and defined in [1]. Provide a reward function specification in LTL for the following office world task: the agent is required to get to a state where coffee is true, and then back to the office, and after that maintain stop forever, all the while without stepping on decorations. Use coffee, office, stop and decs for propositions.

**W1-2** Describe a path that satisfies $G\,p \rightarrow F q$ and does not satisfy $G\,(p \rightarrow F q)$.

**W1-3** In paper [2], a logic $LTL_f$ is defined, that interprets LTL formulas on finite traces. Let the length of a trace $\lambda$ be $length(\lambda)$ and $last(\lambda) = length(\lambda) - 1$ (counting from 0). The truth definition for $LTL_f$ modalities is as follows:

- $\lambda \models X\varphi$ iff $0 < last(\lambda)$ and $\lambda, [1, \ldots, last(\lambda)] \models \varphi$; <- we are checking for the next element/state
- $\lambda \models \varphi\, U\, \psi$ iff for some $j$ such that $0 \leq j \leq last(\lambda)$, $\lambda[j, \ldots, last(\lambda)] \models \psi$ and $\lambda[k, \ldots, last(\lambda)] \models \varphi$ for all $k$ with $0 \leq k < j$.

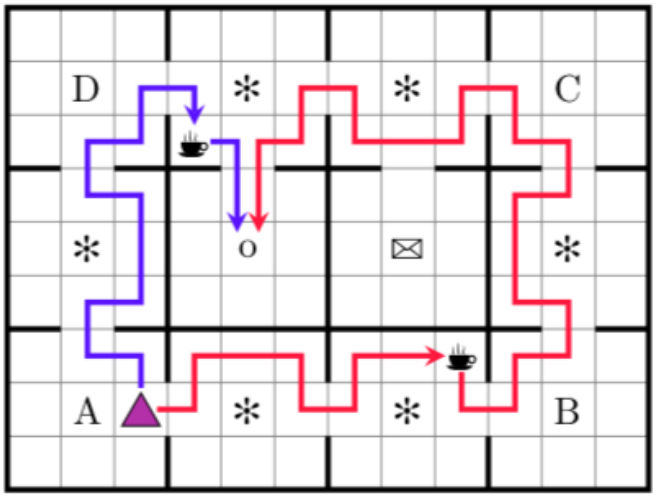


Figure 1: Office World Domain

Give an example formula that is true on some finite trace and false on all infinite traces.

## Tasks that can be done during Week 2 (w/c 3 May)

The following tasks can be done after watching the lectures on CTL and Model Checking with MCMAS and form part of your coursework submission.

**W2-1** The *vacuum world* domain is shown in Figure 2 and introduced in [3]. Create a description in ISPL of the following vacuum cleaner agent: the variables are inA, inB, cleanA, cleanB, and actions are: $\{nil, move, suck\}$, where $nil$ is doing nothing, $move$ moves to another room, and $suck$ vacuums the room where the vacuum cleaner is (making the room clean). The initial state is when the vacuum cleaner is in room A and both rooms are dirty.

The following formulas should be true there:

- $\mathsf{inA} \wedge \neg\mathsf{inB} \wedge \neg\mathsf{cleanA} \wedge \neg\mathsf{cleanB}$
- $\mathsf{inA} \rightarrow EF\mathsf{inB}$
- $\mathsf{inA} \rightarrow EF\mathsf{cleanA}$
- $EF(\mathsf{cleanA} \wedge \mathsf{cleanB})$

Extract a witness for $EF(\mathsf{cleanB} \wedge \mathsf{cleanB})$. and the following formulas false:

- $\mathsf{cleanB}$
- $AG\mathsf{inA}$

- $AF\mathsf{inB}$

Extract a counterexample for $AG\mathsf{inA}$.

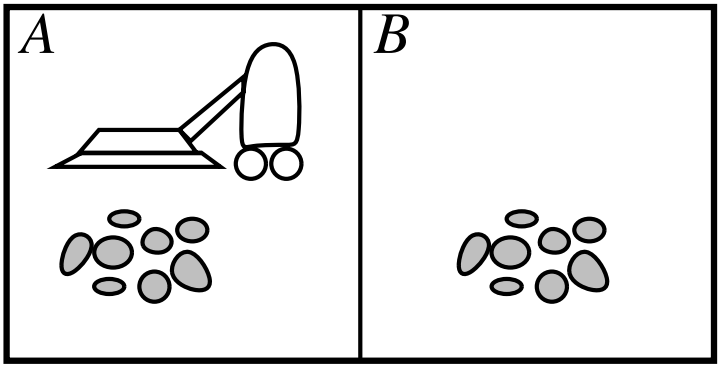


Figure 2: Vacuum World Domain

**W2-2** Given the `one-robot+carriage.ispl` file, translate the following English properties in CTL (you can use ascii encoding as in MCMAS) and check using MCMAS whether they hold. Submit the translations together with the MCMAS outcomes (and witnesses/counterexamples if appropriate).

- it is possible to avoid pos2 forever
- it is possible to be in position pos2 in the next step
- it is possible in the future to be in position pos1 and in the next step after that in pos2
- it is possible to be in pos0 until reaching pos2
- it is possible to be in pos0 or pos1 until reaching pos2
- it is possible to always have pos1 reachable in at most 2 steps

**W2-3** The CTL truth definition counts the present state as 'future': for example, $M, q \models E\varphi\, U\, \psi$ if $M, q \models \psi$. Consider a definition that instead does not count the present as the future:
$M, q \models E\varphi\, U'\psi$ iff there exists a path $\lambda$ from $q$ such that for some $i > 0$, $M, \lambda[i] \models \psi$, and for all $j$ with $0 \leq j < i$, $M, \lambda[j] \models \varphi$.

Write the case for the model checking algorithm for CTL with this $EU'$ modality.

# References

[1] Alberto Camacho, Rodrigo Toro Icarte, Toryn Q. Klassen, Richard Anthony Valenzano, and Sheila A. McIlraith. LTL and beyond: Formal languages for reward

function specification in reinforcement learning. In Sarit Kraus, editor, *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019*, pages 6065–6073. ijcai.org, 2019. https://doi.org/10.24963/ijcai.2019/840.

[2] Giuseppe De Giacomo and Moshe Y. Vardi. Linear temporal logic and linear dynamic logic on finite traces. In Francesca Rossi, editor, *IJCAI 2013, Proceedings of the 23rd International Joint Conference on Artificial Intelligence, Beijing, China, August 3-9, 2013*, pages 854–860. IJCAI/AAAI, 2013. http://www.aaai.org/ocs/index.php/IJCAI/IJCAI13/paper/view/6997.

[3] Stuart Russell and Peter Norvig. *Artificial Intelligence: A Modern Approach*. 1995. http://aima.cs.berkeley.edu.