	Chương 1	Chương 2	Chương 3
Lê Huỳnh Trúc Vy	**1-4**	**1-3**	**1, 2**
Trần Mỹ Yên	**5-8**	**4,5,8**	**3, 4**
Thái Minh Khang	**9-12**	**6,7**	**5, 6, 7**
Lê Quốc Thái	**13-16**	**9-11**	**8, 9, 10**

Hạn chót: Thứ 5 (18/9)

CHƯƠNG 1

1. Goal/reward alignment: How do we specify a robust objective function? Whose objectives are used?

2. Instrumental convergence: All intelligent agents will pursue common subgoals like the need for more power to get better at reaching its objectives. How will this need be balanced with human's needs?

3. What are the LLMs? Percepts? Actions? Objectives?

4. How do Large Language Models fit into the AI Framework in this Course? think like a human? act like a human? think rationally? act rationally?

5. What do LLMs do? Do LLMs act rationally? Ask a chatbot if it • acts rational

• Is an intelligent agent

6. Would a modern LLM pass the Turing Test?

• Would you be fooled?

• Why does it or does it not pass your test?

• What does this mean for artificial general intelligence (AGI) or narrow AI?

7. How do we currently test the performance of LLMs?

8. How do you think LLMs will affect the value of being able to write assays as taught in high school?

9. LLMs write computer code. What does this mean for the value of learning to code?

10. When should students be allowed to use the following tools? Give reasons for your decision.

- A pocket calculator
- LLMs (to answer homework questions and write assays)
- LLMs to write or support writing code

11.How are LLMs affected by:

- Robustness: Black swan vs. adversarial robustness

- Monitoring AI
- What about liability?
- Goal/reward alignment
- Reward hacking
- AGI and instrumental convergence

12. Should the use of LLMs be regulated? How? What about copyright?
13. How do LLMs reason and what are the limits?
14. How do we make sure that LLMs generate factually correct output?
15. How do we fairly compensate the people who create the data that is used to train LLMs?
16. How do we use LLMs in learning, so human learning is not compromised?

CHƯƠNG 2

1. If we have two cars and one provides more (expected) utility. Which car is rational?
2. Can a rational self-driving car be involved in an accident?
3. How would a self-driving car explore and learn?
4. What does bounded rationality mean for a self-driving car?

5.

Environment for a Self-Driving Car

☐ **Fully observable:** The agent's sensors always show the whole **state**.

vs.

☐ **Partially observable:** The agent only perceives part of the **state** and needs to remember or infer the test.

Deterministic:

☐ **a)** **Percepts** are 100% reliable

☐ b) Changes in the environment are completely determined by the current **state** of the environment and the agent's **action**.

vs.

Stochastic:

☐ **a)** **Percepts** are unreliable (noise distribution, sensor failure probability, etc.). This is called a stochastic sensor model.

☐ b) The **transition function** is stochastic leading to transition probabilities and a Markov process.

☐ **Known:** The agent knows the **transition function**.

vs.

☐ **Unknown:** The needs to **learn the transition function** by trying actions.

☑ Check what applies and explain what it means for a self-driving car.

6. Example: Solving a puzzle. What action gets me closer to the solution?

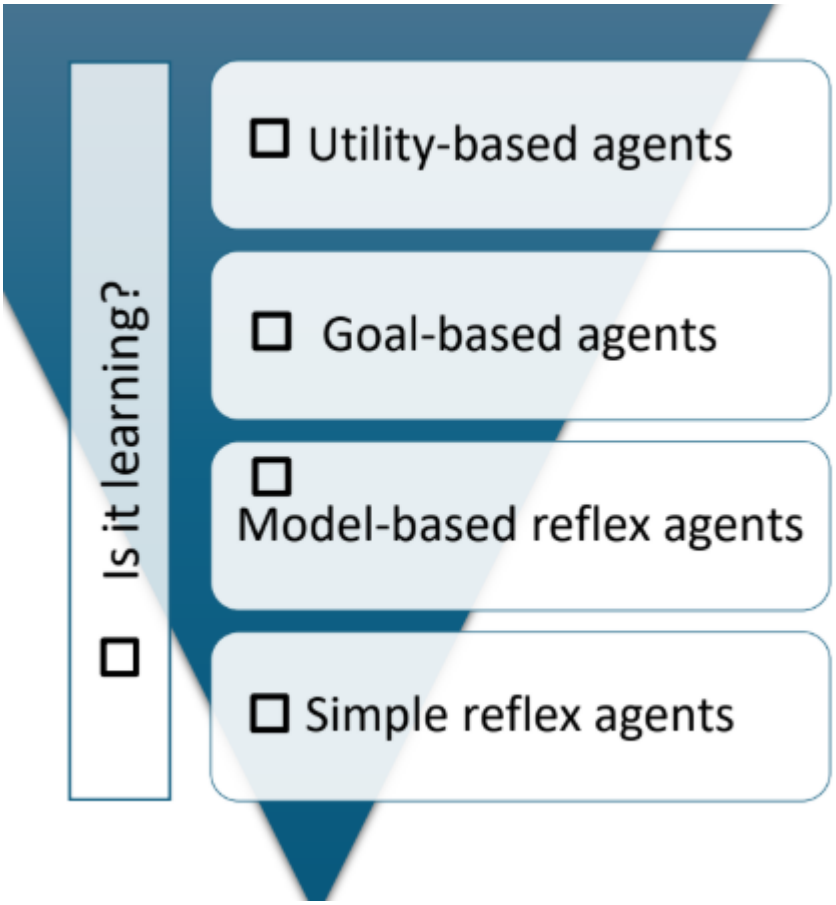
7. Design a structured representation for the state of a self-driving car.

   a) What fluents should it contain?

   b) What actions can cause transitions?

   c) Draw a small transition diagram.

8. What Type of Intelligent Agent is a Self-Driving Car?



9. Does it collect utility over time? How would the utility for each state be defined?

10. Does it have a goal state?

11. Does it store state information. How would they be defined (atomic/factored)?

12. Does it use simple rules based on the current percepts?

CHƯƠNG 3

1. How do we find the optimal solution (sequence of actions/states)?
2. In how many ways can we order/arrange n objects?
3. What is the State Space Size?
4. What is the Search Complexity?
5. What relaxations are used in these two cases?
6. What is the cost that needs to be estimated?

7. What would be a heuristic value for these boards:

8. How do you calculate the heuristic value?
9. Is the heuristic admissible?

10. Does the heuristic use relaxation?