**TRƯỜNG ĐẠI HỌC SÀI GÒN**

KHOA CÔNG NGHỆ THÔNG TIN



A blue circle with white text

Description automatically generated

**EXERCISE 1: TRẢ LỜI CÂU HỎI TRONG CHAPTER 1, 2, 3 (nhóm)**

HỌC PHẦN: TRÍ TUỆ NHÂN TẠO NÂNG CAO

**Giảng viên hướng dẫn** : Ts. Đỗ Như Tài

**Nhóm 11:**

Phạm Văn Nam 3122410251

Nguyễn Quan Tuấn Nghĩa 3122410260

Tạ Hồng Quí 3122410348

Vũ Quốc Vương 3120410629

*Thành phố Hồ Chí Minh - Tháng 09/2025*

Mục lục

[Bảng Phân Công 3](#_Toc208915724)

[Chapter 1: Introduction Discusion 3](#_Toc208915725)

[Discussion 1: The Goal of AI 3](#_Toc208915726)

[Discussion 2: AI Safety and Optimizers 4](#_Toc208915727)

[Discussion 3: Large Language Models (LLMs) 4](#_Toc208915728)

[Discussion 4: Turing Test: Large Language Models 4](#_Toc208915729)

[Discussion 5: The AI Effect: AI gets no respect? 5](#_Toc208915730)

[Discussion 6: AI Safety 5](#_Toc208915731)

[Discussion 7: Conclusion 5](#_Toc208915732)

[Chapter 2: Intelligent Agents Discusion 5](#_Toc208915733)

[Discussion 1: Self-driving Cars 5](#_Toc208915734)

[Discussion 2: PEAS Description of the Environment of a Self-Driving Car 6](#_Toc208915735)

[Discussion 3: Environment for a Self-Driving Car 6](#_Toc208915736)

[Discussion 4: State Representation: Self-Driving Car 6](#_Toc208915737)

[Discussion 5: What Type of Intelligent Agent is a Self-Driving Car? 7](#_Toc208915738)

[Chapter 3: Solving problems by searching 7](#_Toc208915739)

[Discussion 1: Heuristics from Relaxed Problems 7](#_Toc208915740)

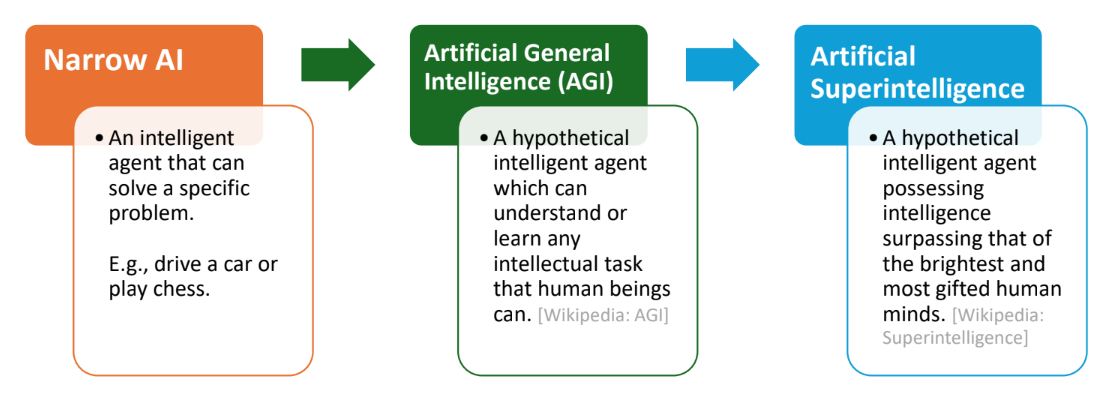
[Discussion 2: Case Study: Heuristic for Tic-Tac-Toe 7](#_Toc208915741)

# Bảng Phân Công

|  |  |  |  |
| --- | --- | --- | --- |
| **STT** | **Công việc** | **Người thực hiện** | **Trạng thái** |
|  |  | Văn Nam |  |
|  |  | Tuấn Nghĩa |  |
|  |  | Hồng Quí |  |
|  |  | Quốc Vương |  |
| **Đánh giá**: | | | |

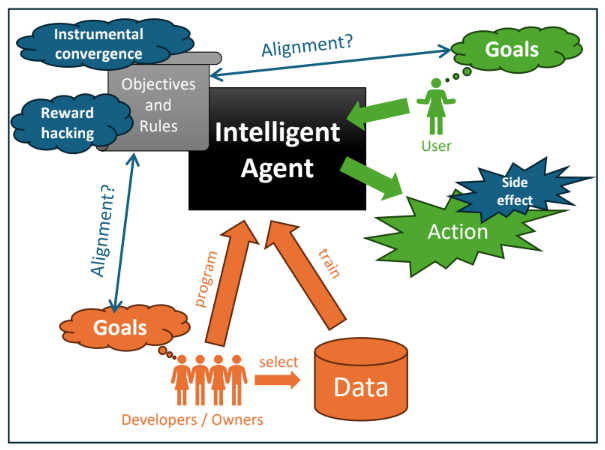
# Chapter 1: Introduction Discusion

## Discussion 1: The Goal of AI



* How can we achieve this? Create an agent that can:
  + Think like a human?
  + Act like a human?
  + Think rationally?
  + Act rationally?

## Discussion 2: AI Safety and Optimizers



* Goal/reward alignment: How do we specify a robust objective function? Whose objectives are used?
* Reward hacking: The Al learns to exploit unintended side effects to get a high "score" without solving the objective. Al needs to follow social norms.
* 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?

## Discussion 3: Large Language Models (LLMs)

* How do Large Language Models fit into the Al Framework in this Course?
* What do LLMs do?
* Do LLMs act rationally?

## Discussion 4: Turing Test: Large Language Models

* 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?
* How do we currently test the performance of LLMs?
* See: Open LLM Leaderboard (Hugging Face)

## Discussion 5: The AI Effect: AI gets no respect?

* How do you think LLMs will affect the value of being able to write assays as taught in high school?
* LLMS write computer code. What does this mean for the value of learning to code?
* 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

## Discussion 6: AI Safety

* How are LLMs affected by:
* Robustness: Black swan vs. adversarial robustness
* Monitoring Al
* What about liability?
* Goal/reward alignment
* Reward hacking
* AGI and instrumental convergence
* Should the use of LLMs be regulated?
* How?
* What about copyright?

## Discussion 7: Conclusion

* How do LLMs reason and what are the limits?
* How do we make sure that LLMs generate factually correct output?
* How do we fairly compensate the people who create the data that is used to train LLMs?
* How do we use LLMs in learning, so human learning is not compromised?

# Chapter 2: Intelligent Agents Discusion

## Discussion 1: Self-driving Cars

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

## Discussion 2: PEAS Description of the Environment of a Self-Driving Car

The PEAS description.

|  |  |  |  |
| --- | --- | --- | --- |
| Performance measure | Environment | Actuators | Sensors |
|  |  |  |  |

## Discussion 3: Environment for a Self-Driving Car

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

|  |  |
| --- | --- |
| Fully observable: The agent's sensors always show the whole state. | Partially observable: The agent only  perceives part of the state and needs to remember or infer the test. |
| Deterministic | Stochastic |
| a) Percepts are 100% reliable. | a) Percepts are unreliable (noise distribution, sensor failure probability, etc.). This is called a stochastic sensor model. |
| b) Changes in the environment are completely determined by the current state of the environment and the agent's action. | b) The transition function is stochastic leading to transition probabilities and a Markov process. |
| Known: The agent knows the transition  function. | Unknown: The needs to learn the transition function by trying actions. |

## Discussion 4: State Representation: Self-Driving Car

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

* What fluents should it contain?
* What actions can cause transitions?
* Draw a small transition diagram.

## Discussion 5: What Type of Intelligent Agent is a Self-Driving Car?

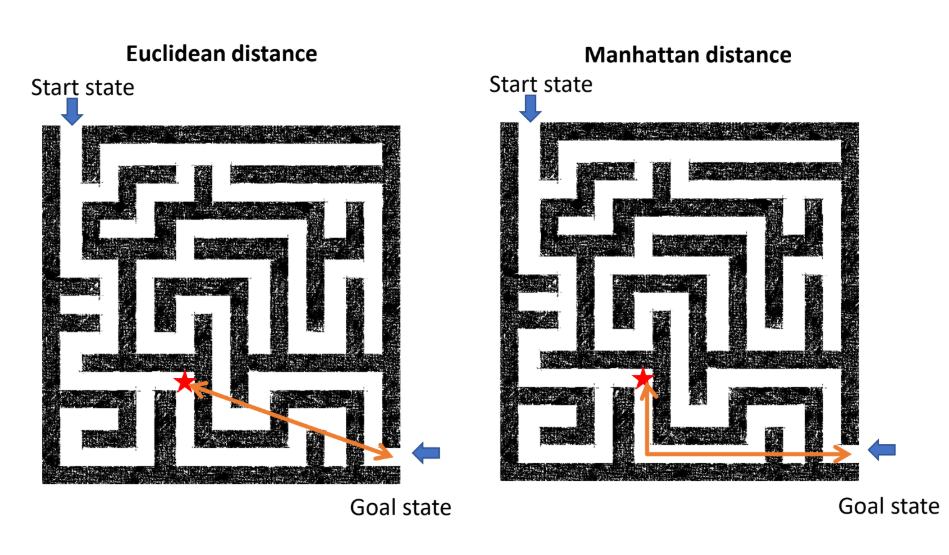
Check what applies

|  |  |
| --- | --- |
| Utility-based agents | Does it collect utility over time? How would the utility for each state be defined? |
| Goal-based agents | Does it have a goal state? |
| Model-based reflex agents | Does it store state information. How would they be defined (atomic/factored)? |
| Simple reflex agents | Does it use simple rules based on the current percepts? |

# Chapter 3: Solving problems by searching

## Discussion 1: Heuristics from Relaxed Problems

What relaxations are used in these two cases?



## Discussion 2: Case Study: Heuristic for Tic-Tac-Toe

* Define the goal states:
* What is the cost that needs to be estimated?
* What would be a heuristic value for these boards:



* How do you calculate the heuristic value?
* Is the heuristic admissible?
* Does the heuristic use a relaxation?