

# NHÓM 06

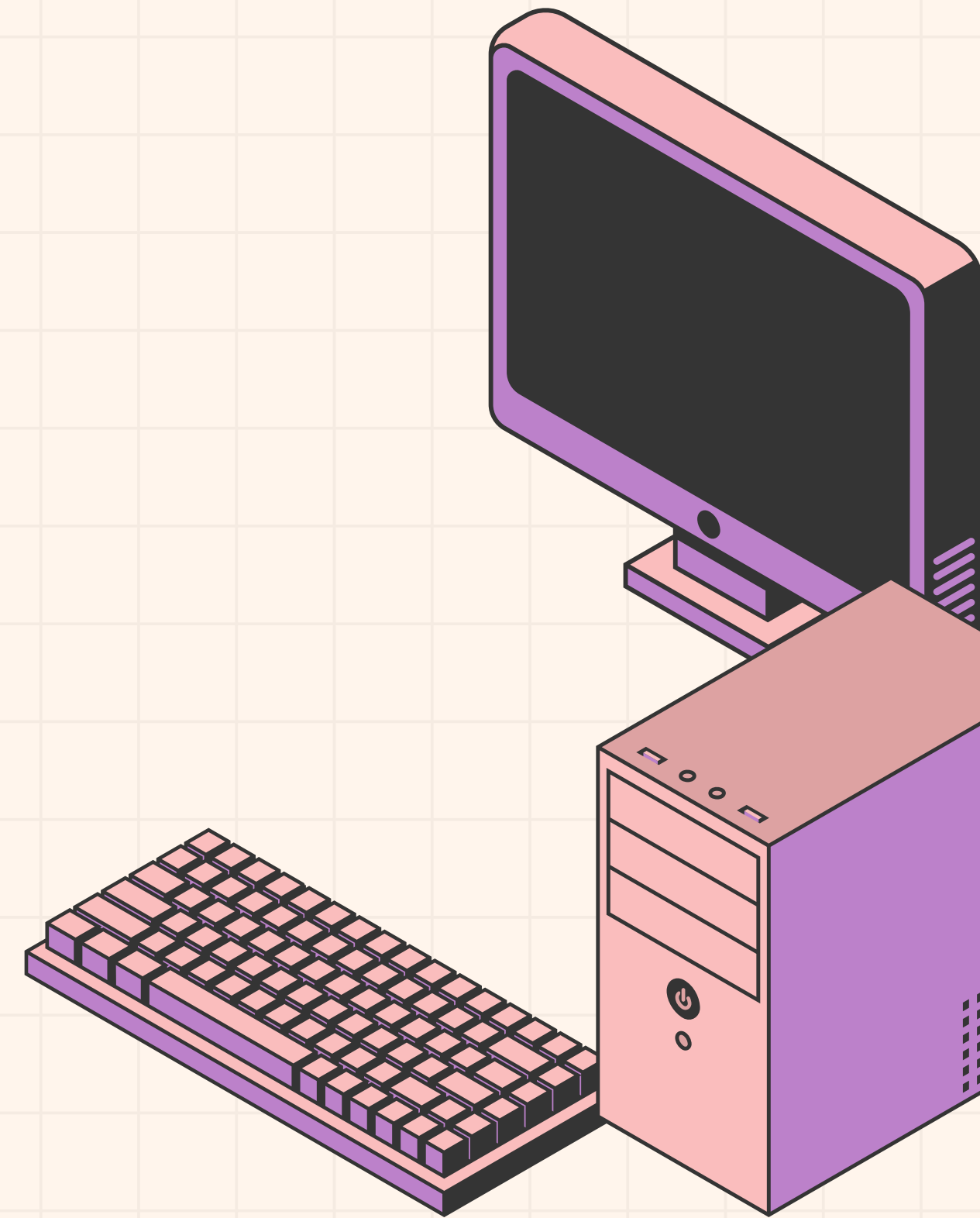
THÀNH VIÊN:

NGUYỄN VĂN ANH TÚ - 3122410445

TRƯƠNG HỮU NGHĨA - 3122410263

ĐỖ KHÔI NGUYỄN - 3122410266

LAI TẤN TÀI - 3122410366



# CHƯƠNG 1

Percepts: các dữ liệu đầu vào mà nó nhận được từ người dùng. Trong ví dụ này, các cảm nhận là hai lời nhắc (prompts) của người dùng

Actions: là các phản hồi mà nó tạo ra để đáp lại cảm nhận. Trong ví dụ này, hành động của LLM là:

1.Đưa ra câu trả lời đồng ý giúp đỡ và yêu cầu thông tin chi tiết hơn: "Of course! As an AI language model, I would be happy to assist you..." (Tất nhiên! Là một mô hình ngôn ngữ AI, tôi rất vui lòng hỗ trợ bạn...).

2.Tạo ra một bài luận dài 1000 từ về các tác phẩm của Shakespeare, bắt đầu bằng "Certainly! Here is a 1000-word essay on the works of Shakespeare..." (Chắc chắn rồi! Đây là một bài luận 1000 từ về các tác phẩm của Shakespeare...).

Ojectives: là phản hồi các lời nhắc của người dùng một cách hiệu quả và chính xác nhất có thể. Cụ thể hơn, mục tiêu của nó là:


1.Hỗ trợ người dùng: Cung cấp thông tin, hướng dẫn hoặc tạo ra nội dung theo yêu cầu.

2.Hoàn thành tác vụ: Tạo ra một bài luận 1000 từ về một chủ đề cụ thể.


3.Duy trì cuộc hội thoại: Phản hồi theo một cách tự nhiên và mạch lạc để mô phỏng một cuộc trò chuyện với con người.

## Example Conversation With a Large Language Model (LLM)

 Can you help me with my english homework?

 Of course! As an AI language model, I would be happy to assist you with your English homework. Just let me know what you need help with, and I will do my best to provide you with the information and guidance you need.

 Please write an 1000 word essay on the works of Shakespear.

 Certainly! Here is a 1000-word essay on the works of Shakespeare:

William Shakespeare is widely regarded as one of the greatest writers in the history of the English language. Born in Stratford-upon-Avon, England, in 1564, Shakespeare began his career as an actor before eventually becoming a playwright. He wrote plays, sonnets, and other works that have been studied, performed, and celebrated for centuries.

Shakespeare's plays are known for their rich characters, intricate plots, and poetic language. His works have been categorized into three main genres: tragedies, comedies, and histories. Each of these genres explores different themes and motifs, but they all share Shakespeare's distinct voice and style. ...

Source: <https://chat.openai.com/>

What are the LLMs

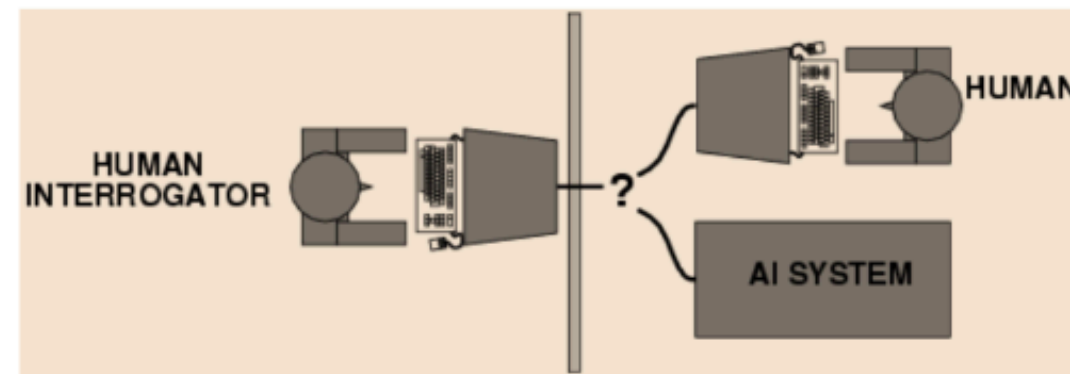
Percepts?

Actions?

Objectives?

# CHƯƠNG 1

## Turing Test: Large Language Models (LLMs)



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 \(Hugging Face\)](#)

act like a human?

How do LLMs relate to this:

### Chinese Room Argument



Thought experiment by John Searle (1980): Imitate intelligence using rules.

# CHƯƠNG 1

Câu hỏi: Would a modern LLM pass a 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?

Trả lời dựa trên bài báo tham khảo ở arxiv.org thì

- Các tác giả đã đánh giá bốn hệ thống (ELIZA, GPT-4o, LLaMa-3.1-405B và GPT-4.5) trong hai bài kiểm tra Turing được kiểm soát.
- Những người tham gia trò chuyện 5 phút cùng lúc với một người và một trong các hệ thống AI này, sau đó đánh giá xem ai là con người.
- Khi được yêu cầu sử dụng một "nhân vật giống con người" (humanlike persona), GPT-4.5 đã được đánh giá là con người tới 73% số lần, cao hơn đáng kể so với tỷ lệ người tham gia thực sự được xác định là người thật.
- LLaMa-3.1 được đánh giá là con người 56% số lần, trong khi các mô hình cơ bản như ELIZA và GPT-4o đạt tỷ lệ thấp dưới mức ngẫu nhiên.
- Các tác giả khẳng định rằng kết quả này là bằng chứng thực nghiệm đầu tiên cho thấy một hệ thống nhân tạo đã vượt qua bài kiểm tra Turing ba bên tiêu chuẩn.
- Tuy nhiên, bài báo cũng nhấn mạnh rằng việc vượt qua bài kiểm tra Turing không có nghĩa là AI đạt đến "trí thông minh cấp độ con người", mà chỉ cho thấy khả năng "bắt chước" rất khéo léo.



# CHƯƠNG 1



## The AI Effect: AI gets no respect?

---

- How do you think LLMs will affect the value of being able to write **essays** 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 essays)
  - LLMs to write or support writing code

# CHƯƠNG 1

- Về LLM: Mô hình ngôn ngữ lớn (LLM) là một AI có khả năng xử lý và tạo văn bản giống con người. Dữ liệu đầu vào của nó là các yêu cầu của người dùng (percepts), và đầu ra là các phản hồi hoặc nội dung được tạo ra (actions). Mục tiêu của LLM là thực hiện các tác vụ ngôn ngữ một cách hiệu quả và tự nhiên.
- Về ảnh hưởng của LLM đến giáo dục:
- Viết luận: LLM làm giảm giá trị của kỹ năng viết luận cơ bản, nhưng lại nâng cao tầm quan trọng của tư duy phản biện và sáng tạo. Học sinh cần học cách sử dụng LLM như một công cụ hỗ trợ chứ không phải công cụ gian lận.
- Viết code: LLM giúp lập trình viên tiết kiệm thời gian cho các công việc lặp lại, cho phép họ tập trung vào các vấn đề phức tạp hơn. Việc học lập trình sẽ chuyển từ "viết code" sang "sử dụng code do AI tạo ra và giải quyết vấn đề".
- Sử dụng công cụ: Việc cho phép học sinh dùng các công cụ như máy tính bỏ túi hay LLM cần có sự hướng dẫn rõ ràng từ giáo viên, đảm bảo các em vẫn nắm vững kiến thức nền tảng và không lạm dụng công nghệ.

# CHƯƠNG 1

[Administration](#)[Priorities](#)[The Record](#)

OCTOBER 30, 2023

## Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence

 BRIEFING ROOM PRESIDENTIAL ACTIONS

By the authority vested in me as President by the Constitution and the laws of the United States of America, it is hereby ordered as follows:

Section 1. Purpose. Artificial intelligence (AI) holds extraordinary potential for both promise and peril. Responsible AI use has the potential to help solve urgent challenges while making our world more prosperous, productive, innovative, and secure. At the same time, irresponsible use could exacerbate societal harms such as fraud, discrimination, bias, and disinformation; displace and disempower workers; stifle competition; and pose risks to national security. Harnessing AI for good and realizing its myriad benefits requires mitigating its substantial risks. This endeavor demands a society-wide effort that includes government, the private sector, academia, and civil society.

Some important points:

- Artificial Intelligence must be **safe and secure**.
- Promoting **responsible innovation, competition, and collaboration**
- Americans' **privacy and civil liberties** must be protected.

- Should the use of LLMs be regulated?
- How?
- What about copyright?

# CHƯƠNG 1

## 1. Có nên điều chỉnh việc sử dụng LLM không?

- Có, Lệnh Hành pháp này cho thấy việc sử dụng LLM cần được điều chỉnh. Lệnh nhấn mạnh rằng AI nói chung, bao gồm cả LLM, mang lại cả "lời hứa và hiểm họa." Việc sử dụng AI thiếu trách nhiệm có thể gây ra những tổn hại xã hội đáng kể như lừa đảo, phân biệt đối xử, sai lệch thông tin và rủi ro an ninh quốc gia. Do đó, chính phủ cần đóng vai trò trong việc quản lý để đảm bảo sự phát triển và sử dụng AI một cách an toàn, có trách nhiệm và đáng tin cậy.

## 2. Điều chỉnh như thế nào?

Lệnh Hành pháp đề xuất một số cách tiếp cận để điều chỉnh LLM:

- Tiêu chuẩn và Quy trình đánh giá: Yêu cầu các công ty phát triển các mô hình nền tảng có nguy cơ cao phải thông báo cho chính phủ khi đào tạo mô hình và chia sẻ kết quả đánh giá an toàn.
- Phát hiện nội dung do AI tạo ra: Đặt ra các tiêu chuẩn và phương pháp hay nhất để phát hiện nội dung do AI tạo ra và xác thực nội dung chính thức của con người. Điều này có thể bao gồm việc phát triển hệ thống watermarking để dán nhãn rõ ràng nội dung do AI tạo ra.
- Bảo vệ dân quyền và quyền riêng tư: Yêu cầu các cơ quan liên bang xây dựng hướng dẫn để ngăn chặn phân biệt đối xử thuật toán và bảo vệ quyền riêng tư.
- Thúc đẩy đổi mới có trách nhiệm: Đảm bảo rằng các quy định không cản trở sự cạnh tranh và đổi mới trong ngành AI.



# CHƯƠNG 1

## 3. Về bản quyền?

Vấn đề bản quyền được Lệnh Hành pháp đề cập, mặc dù không đi vào chi tiết cụ thể. Lệnh này nêu bật sự cần thiết phải giải quyết các câu hỏi về sở hữu trí tuệ phát sinh từ việc sử dụng các mô hình AI tạo sinh. Điều này ngụ ý rằng các chính sách và quy định mới sẽ cần được phát triển để bảo vệ quyền của người sáng tạo và người phát minh, đồng thời thúc đẩy một hệ sinh thái AI công bằng và cạnh tranh.

# CHƯƠNG 1

## Conclusion

- LLMs are a powerful new generative AI technology which many applications.
- Unfortunately, there are many open questions. For example:
  - 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?



# CHƯƠNG 1

Các mô hình ngôn ngữ lớn (LLMs) đúng là một bước tiến lớn trong AI, nhưng chúng vẫn còn nhiều hạn chế và câu hỏi lớn chưa được giải quyết. Cụ thể:

- Về cách suy luận: LLMs không "suy luận" theo cách con người hiểu. Chúng dự đoán từ dựa trên thống kê từ dữ liệu huấn luyện, chứ không hiểu ý nghĩa thật sự. Giới hạn lớn là chúng có thể "tự tin" đưa ra thông tin sai hoặc không có cơ sở.
- Về độ chính xác: Để đảm bảo đầu ra chính xác, cần kết hợp nhiều biện pháp như kiểm tra thực tế, cập nhật dữ liệu, sử dụng nguồn tham khảo đáng tin cậy, và có con người giám sát.
- Về bồi thường dữ liệu: Đây là vấn đề đạo đức và pháp lý lớn. Cần có khung pháp lý rõ ràng để đảm bảo những người tạo nội dung gốc được ghi nhận và bồi thường xứng đáng, ví dụ như chia sẻ doanh thu hoặc cơ chế cấp phép dữ liệu.
- Về học tập: LLMs có thể hỗ trợ học tập nếu dùng đúng cách, ví dụ như giúp giải thích khái niệm, nhưng cũng có thể làm học sinh lười suy nghĩ. Giải pháp là dạy người dùng cách dùng AI như một công cụ hỗ trợ, không phải thay thế, và luôn khuyến khích tư duy phản biện.

# CHƯƠNG 2

**If we have two cars and one provides more (expected) utility. Which car is rational?**

**In this case, both cars are rational. A rational agent chooses the action to maximise the expected utility. Because both cars perform the best action, they are both rational.**

**Can a rational self-driving car be involved in an accident?**

**A rational self-driving car may be involved in an accident. The rational agent is not a perfect agent, so it can make mistakes and perform unexpected actions. A rational agent is not omniscience, mistakes from sensors can lead to unexpected results.**



# CHƯƠNG 2

## How would a self-driving car explore and learn?

A self-driving car would:

- Explore in simulations, where it can safely test many scenarios.
- Use reinforcement learning, trying different driving policies and receiving rewards (safe driving) or penalties (collisions).
- Adapt in the real world cautiously, mainly exploiting what it has already learned while making small adjustments.

## What does bounded rationality mean for a self-driving car?

Bounded rationality means the car cannot compute the perfect optimal decision because of limited time, computational resources, and incomplete sensor data. Instead, it makes the best possible decision within those limits. For example, instead of evaluating all possible routes, it only considers a limited set of safe and feasible options.

# CHƯƠNG 2

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

## **Fluents**

- **Position & motion:**
  - **Current lane, GPS coordinates, heading, speed, acceleration.**
- **Environment:**
  - **Traffic light status, road type (highway, city street), speed limit.**
  - **Distance to nearby vehicles (front, behind, sides).**
  - **Pedestrian presence, obstacles, weather, road surface condition.**
- **Car's internal state:**
  - **Fuel/battery level.**
  - **Sensor health status.**
  - **Current driving mode (manual/auto, cruise control, etc.).**

# CHƯƠNG 2

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

## **Actions that can cause transitions**

- **Driving controls:**
  - Accelerate, brake, steer left/right, change lane.
- **Signals:**
  - Turn signal, horn, headlights.
- **Navigation actions:**
  - Take exit, follow route, stop at destination.

# CHƯƠNG 2

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

## Small transition diagram

- State 1: [speed = 50, traffic\_light = green] –(light turns yellow)---->
- State 2: [speed = 50, traffic\_light = yellow] --(agent chooses "brake")-->
- State 3: [speed = 0, traffic\_light = red] --(light turns green + agent "accelerate")-->...



# CHƯƠNG 2

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



☐ Is it learning?

- ☐ Utility-based agents
- ☐ Goal-based agents
- ☐ Model-based reflex agents
- ☐ Simple reflex agents

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

Does it have a goal state?

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

Does it use simple rules based on the current percepts?

☒ Check what applies

# CHƯƠNG 2

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

**Yes, A self-driving car continuously evaluates the utility of being in different states (safe vs. risky, smooth vs. uncomfortable, efficient vs. wasteful)**

**Does it have a goal state?**

**Yes. The car's main goal is to transport passengers from point A to point B safely and efficiently.**

# CHƯƠNG 2

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

**Yes.**

**It must maintain internal state, because the environment is partially observable.**

- Example: A car hidden behind a truck is remembered**

**Factored representation: variables like [speed = 60 km/h, lane = 2, traffic\_light = red, car\_ahead\_distance = 15m].**

**Does it use simple rules based on the current percepts?**

**Partially. However, a car cannot rely only on simple rules, since driving requires prediction, planning, and reasoning.**

# CHƯƠNG 3

Examples: What is the Search Complexity?

*b*: maximum branching factor  
*m*: max. depth of tree  
*d*: depth of the optimal solution

Often a rough upper limit is sufficient to determine how hard the search problem is.

Maze	8-queens problem	8-puzzle problem	Tic-tac-toe
<i>b</i> = 4 actions <i>m</i> = longest path to the goal or a dead end (bounded by $x \times y$ ) <i>d</i> = shortest path to the goal (bounded by $x \times y$ )	<i>b</i> = ? What are the actions? Moving one Queen: $64 - 7 = 57$ <i>m</i> = We may have to try all: $\binom{64}{8} \approx 4.4 \times 10^9$ <i>d</i> = move each queen in the right spot = 8	<i>b</i> = 4 actions to move the empty tile. <i>m</i> = Try all $O(9!)$ <i>d</i> = ???	<i>b</i> = 9 actions for the first move. <i>m</i> = 9 <i>d</i> = 9 (if both play optimal)

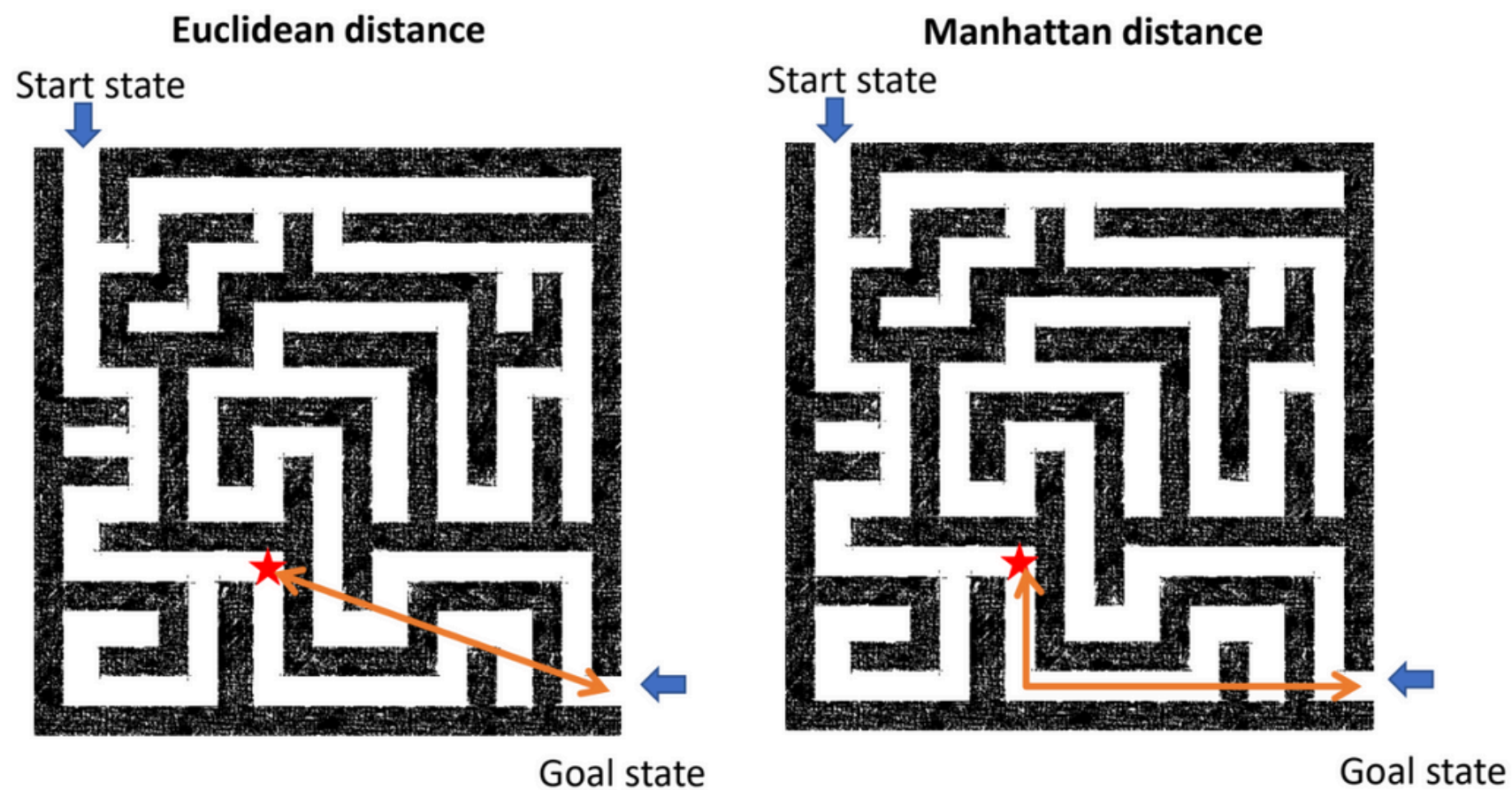
Giá trị *d* của bài toán 8-puzzle problem là :  $x$  , với  $x$  thuộc  $[1,9!]$



# CHƯƠNG 3

## Heuristics from Relaxed Problems

What relaxations are used in these two cases?



## Đối với Euclidean distance:

Cho phép di chuyển theo bất kỳ hướng nào, nếu không sử dụng phương pháp này các di chuyển thường bị ràng buộc bởi các ô kề nhau. Điều này làm giảm nhẹ bài toán bằng cách bỏ qua các ràng buộc về di chuyển trong lưới.

## Đối với Manhattan distance:

Cho phép di chuyển tới trạng thái đích mà bỏ qua các vật cản theo các hướng ngang/dọc mà không bị cản trở

# CHƯƠNG 3

## 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?



– Trạng thái đích: Là trạng thái mà có 1 hàng hoặc 1 cột hoặc 1 đường chéo là x hoặc o hay khi không còn ô trống nào trên bàn cờ.

– Chi phí ước tính để chiến thắng hoặc tránh thất bại là hàm heuristic đánh giá trạng thái bàn cờ.

– Với 2 bảng này, giá trị heuristic sẽ là 0.

– Với thuật toán Minimax, giá trị heuristic được tính là 1 nếu x thắng, -1 nếu o thắng và 0 nếu chưa có ai thắng.

# CHƯƠNG 3

## 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:

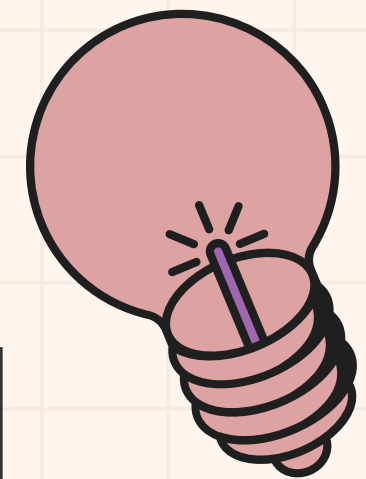
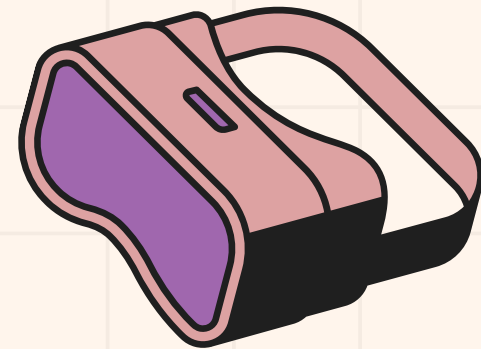


- How do you calculate the heuristic value?
- Is the heuristic admissible?
- Does the heuristic use a relaxation?



– Với hàm heuristic này, thì không được xem là admissible vì hàm heuristic này chỉ đánh giá kết quả cuối cùng của trạng thái.

– Hàm heuristic này không sử dụng “relaxation”. . Mà chỉ đơn giản là hàm đánh giá kết quả cuối cùng của trạng thái.



# THANK YOU

