Assignment 1

TDT4136 - Introduction to Artificial Intelligence
Fall 2023

Alessandro Crespi
Francesca Grimaldi

1 What is Artificial Intelligence (AI)? Include at least 3 definitions of AI that are not covered in the lecture.

The term Artificial Intelligence (AI) was coined in 1955 by emeritus Stanford Professor John McCarthy, who defined it as *"the science and engineering of making intelligent machines, especially intelligent computer programs"* (McCarthy, 2004).

Since then, the field of AI has advanced significantly, and other definitions of the term have been presented that either expand the aforementioned one or place greater emphasis on its details and areas of application.

Dan Patterson (1990, p. 2) stated that *"AI is a branch of computer science concerned with the study and creation of computer systems that exhibit some form of intelligence: systems that learn new concepts and tasks, systems that can reason and draw useful conclusions about the world around us, systems that can understand a natural language or perceive and comprehend a visual scene, and systems that perform other types of feats that require human types of intelligence."*

According to American mathematician Richard Bellman (1978), AI as *"the automation of activities associated with human thinking, which include decision making, problem solving, and learning."*

Governments have also provided definitions of AI in the context of legislation, when attempting to establish and set ethical guidelines on the subject matter. For example, the European Committee's one (2019, p. 1) is that *"Artificial intelligence (AI) refers to systems that display intelligent behavior by analyzing their environment and taking actions – with some degree of autonomy – to achieve specific goals."*

2 What is the Turing test? What is its purpose and how is it conducted? Are there any new proposals for the Turing Test?

The Turing test is a behavioral assessment method designed to evaluate a machine's (or program's) capacity to exhibit intelligent behavior that is equivalent (at least partially) to that of a human being.

Turing's original proposal was the following: the machine being evaluated would be put against a human, and both would have to respond to the questions provided by an interrogator. When asking the questions, the interrogator is not aware of which of the interlocutors they are interacting with; they must determine, from the response they receive, whether they are speaking to the machine or to the person. The machine passes the test if it manages to fool the interrogator for at least 30% of the test's time.

This test's main premise is that verbal behavior alone is enough to assess an agent's intelligence level.

Different variations of the original Turing Test have been proposed in an attempt to develop a stronger "benchmark" of intelligence (Total Turing Test, 2010). Here are two of the new proposals:

- **Total Turing Test**
  This variation, proposed by cognitive scientist Stevan Harnad, adds two further requirements to the traditional Turing test (Turing test, 2015, para. 6.5 Total Turing test). Using computer vision and robotics, the questioner may additionally assess the subject's perception and object-manipulation skills.
- **Reverse Turing Test**
  This version of the Turing test is a variant of the original one in which one or more of the roles between machines and humans have been switched around, as the word *reverse* suggests. An example of an implementation is the CAPTCHA (Completely Automated Public Turing test to tell Computers and Humans Apart) mechanism, which many web applications now use to stop automated programs from abusing the site. The user is presented with distorted alphanumerical characters and is asked to type them out (Turing test, 2015, para. 6.1 Reverse Turing test and CAPTCHA), to demonstrate that he/she is a real person. A system that can properly interpret the warped image is most likely to be a human, because software that is sophisticated enough to do so does not exist, or is unavailable to the typical user.

3 What is rationality and what is the difference between thinking rationally and acting rationally? Is rational thinking an absolute condition for acting rationally?

Rationality refers to the ability of an entity to make decisions and take actions that are logically sound and lead to achieving its goals effectively.

The process of deriving conclusions or making decisions based on logical reasoning, formal logic rules and well-known knowledge is referred to as **thinking rationally**.

When **acting rationally**, one does what is expected to maximize goal attainment while taking into account the available information and computational abilities.

Rational thinking is not an absolute condition for acting rationally because the latter does not necessarily involve thinking but may be driven by reflexes or instincts. As an example, blinking eyelids automatically when exposed to bright light is an instinctual response that is triggered by an event without conscious thought.

4 What is the connection between knowledge and action according to Aristotle? How can his argument be used to implement his idea in AI?

Aristotle formulated the concept of **syllogism** (Greek for *with logics*), which theoretically allowed one to generate conclusions mechanically from the initial premises. The following is a famous example to illustrate its basic structure (Syllogism, para. 1):

Major premise: *All men are mortal.*
Minor premise: *Socrates is a man.*
Conclusion: *Socrates is mortal.*

In this case, the conclusion is a speculative proposition.

To highlight the connection between **knowledge** and **action** (an important concern for AI because intelligence requires both action and reasoning), he argued that the conclusion that results from the two premises could also be an action. Starting from the logical link between goals and knowledge of the action's consequences, the rational pattern can produce the action that the entity needs to perform in order to achieve the desired objective.

In the *Nicomachean Ethics*, he offered an algorithm to follow. The key idea is that it is not a matter of ends, but of **means**. He suggests that one should behave in a similar manner to a doctor, who does not question whether he shall heal, but assumes it as his goal and only thinks about how and by what means it can be attained. If it is achievable only by one means, then the problem becomes how to achieve this means, and so on until they come to the first cause. If there are obstacles, one should stop.

In an actual implementation considering his ideas, the AI would attempt to infer the next goal from well-known knowledge, through training and logical schemas based on anticipated outcomes. Then it goes about pursuing the desired goal by considering more possibilities and optimal ways,

choosing a path and continuing looking for the best options for this one, until they reach the first cause.

4.1 Who was (or were) the first AI researcher(s) to implement these ideas?

Newell and Simon were the first AI researchers to put this concept into practice, in 1957.

4.2 What is the name of the program or system they developed? Write a short description about it.

The program that they have developed was called **General Problem Solver** (GPS). By utilizing heuristic search and problem reduction strategies, it was designed to address a wide range of problems. It operated on a formalized representation of problems and used means-ends analysis to break down complex problems into simpler sub-problems, gradually working towards a solution. In order to refine its problem-solving techniques over time, GPS also incorporated aspects of trial-and-error learning.

5 Consider a robot with the task of crossing the road, and an action portfolio A: A = {lookBack, lookForward, lookLeft, lookRight, goForward, goBack, goLeft, goRight}

5.1 While crossing the road, an elk crashes into the robot and smashes it. Is the robot rational?

The robot could be rational considering the possibility that it might have detected the elk and which direction it was taking, but despite taking the necessary measures to avoid the threat (e.g., going back), it was smashed anyway.

It has the possibility to look around (lookBack, lookForward, lookLeft, lookRight), therefore if it does not look or looks but continues in that direction anyway, it is not rational.

5.2 While crossing the road on a green light, a passing car drives into the robot and crashes, preventing the robot from crossing to the other side. Is the robot rational?

Even though the pedestrian traffic light is green and the robot is already moving, it must be designed to check for any eventual dangerous objects coming from the sides, like a car that won't stop by the crosswalk.

If the robot is developed in such a manner and attempts to avoid the car (e.g., by moving in another direction) but cannot prevent the collision because for instance the car is too fast, then is it rational; otherwise, it is not.

6 Consider the vacuum cleaner world described in Figure 2.2 (Chapter 2.1 of AIMA 4th Ed.). Let us modify this vacuum environment such that the agent is penalised with 1 point for each movement:

6.1 Could a simple reflex agent be rational for this environment? Why?

No, the vacuum cleaner in this environment is not rational because it is placed in a partially observable world where it is unable to minimize the movement penalty.

Because of how it is built, it cannot know whether the other square was just cleaned or needs to be cleaned. It has to move and check every time. And after the change of position, there would be the same problem but for the previous square. Therefore, it runs the risk of being stuck in an infinite loop of movements from one square to the other.

6.2 Could a reflex agent with state be rational in this environment? Why?

Yes, it is rational as the cleaner is no longer a simple reflex agent but a more complicated one that has memory of when a specific square was last cleaned. It is now able to minimize the penalty, for instance by switching from one square to another only after a set amount of time.

6.3 Assume now that the simple reflex agent (i.e., with no internal state) can perceive the clean status of both locations at the same time. Could this agent be rational? Why? In case it could be rational, write the agent function using mathematical notation or a table.

Yes, this agent could be rational. Despite having no internal state, the fact that it can perceive both squares (which makes the environment fully observable) means that it could move and suck only when the squares are actually dirty.

An agent function could be the following, in which the parameter `location` indicates the current position of the vacuum cleaner (that can be either A or B), `cleanA` and `cleanB` are booleans whose value is `True` if the square is clean and `False` if it is dirty:

```
def agent_function([location, cleanA, cleanB]):
    if location == A:        #location can be A or B
        if cleanA == False:
            suck()
        else if cleanB == False:
            moveToB()
        else:
            skip()        #stay in the same place and do nothing
    else:
        if cleanB == False:
            suck()
        else if cleanA == False:
            moveToA()
        else:
            skip()        #stay in the same place and do nothing
```

7 Consider the original vacuum cleaner environment shown in Figure 2.2. Describe the environment using the properties from Chapter 2.3.2 (e.g. episodic/sequential, deterministic/stochastic, etc.) Explain why you chose such values and properties.

Partially observable/fully observable
The vacuum cleaner has a sensor that only detects dirt in the location where it is standing; it is unable to tell whether or not there is dirt in other squares, making the environment **partially observable**. It would be fully observable if there was a sensor capable of detecting changes throughout the whole working area.
Single agent/multi-agent
The environment is **single agent**. The performance of the agent is not affected by other agents. There are no competitive neither cooperative interactions.
Deterministic/stochastic
The environment is **deterministic** because its current state and the action executed by the agent completely determine its next state, without uncertainty.
Episodic/sequential
Every action undertaken has no effect on any future environment event and depends only on the state episode itself. For this, the environment is **episodic**.
Dynamic/static
The environment is **dynamic** because it can change while an agent is deliberating. As an example, imagine that the vacuum is on a cleaned square and is deciding to move to the other one, but in

the meantime, the initial square becomes dirty.
Discrete/continuous
The cardinality of the state set, percepts, and actions is **discrete**, and so is the environment.
Known/unknown
We can consider the environment in object as a complete **known** one because there is full knowledge about how it could evolve.

8 Write both advantages and limitations of the following types of agents:

8.1 Simple reflex agents

+ Easy to implement based on the function map of current percepts to actions.
- Will only work if the environment is fully observable.
- May be stuck in infinite loop because of the lack of previous state memory.

8.2 Model-based reflex agents

+ Best solution to turn a partially observable environment into a fully observable one using well-known data and knowledge about the world where the agent is situated.
- It is difficult to exactly determine the current state in partially observable environments.
- May need to guess the current environment state.

8.3 Goal-based agents

+ More flexible since the desired states (goals) are more explicit and various ways of getting closer to the target could be taken into consideration.
- Less efficiency with respect to simpler agents as reflex ones.

8.4 Utility-based agents

+ In the next state during the agent's work cycle, it will be granted the best situation according to the utility function and the performance measurement.
- It is difficult to estimate the utility function based on the current knowledge, as the utility of an action is hard to quantify.
- Require ingenious algorithms.

References

- Bellman, R. (1978). *An Introduction to Artificial Intelligence: Can Computers Think?*. San Francisco: Boyd & Fraser Pub.
- High-Level Expert Group on Artificial Intelligence (European Committee). (2019). *A definition of AI: Main capabilities and scientific disciplines.* Available at: https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai (Accessed: 7 September 2023).
- McCarthy, J. (2004). *What is Artificial Intelligence?*. Stanford University.
- Öztürk, P. (2023). *TDT4136 Introduction to Artificial Intelligence - Lecture 1: Introduction.* [PowerPoint presentation]. NTNU
- Öztürk, P. (2023). *TDT4136 Introduction to Artificial Intelligence - Chapter 2: Intelligent Agents.* [PowerPoint presentation]. NTNU
- Patterson, D. W. (1990). *Introduction to Artificial Intelligence and Expert Systems.* Prentice-Hall International.
- Russell, S. J. & Norvig, P. (2020). *Artificial intelligence: a modern approach.* 4th Global ed. Boston: Pearson.
- Syllogism. [no date]. *New World Encyclopedia.* Available at: https://www.newworldencyclopedia.org/entry/Syllogism (Accessed: 7 September 2023).
- Total Turing Test. (2010). *University of Alberta's Dictionary of Cognitive Science.* Available at: http://www.bcp.psych.ualberta.ca/mike/Pearl_Street/Dictionary/contents/T/totalTuring.html (Accessed: 7 September 2023).
- Turing test. (2015). *Wikipedia.* Available at: https://en.wikipedia.org/wiki/Turing_test. (Accessed: 7 September 2023).