TECHNISCHE UNIVERSITÄT DARMSTADT

Fachbereich Informatik

hessian.AI

AI101

Lecture 2: AI Systems

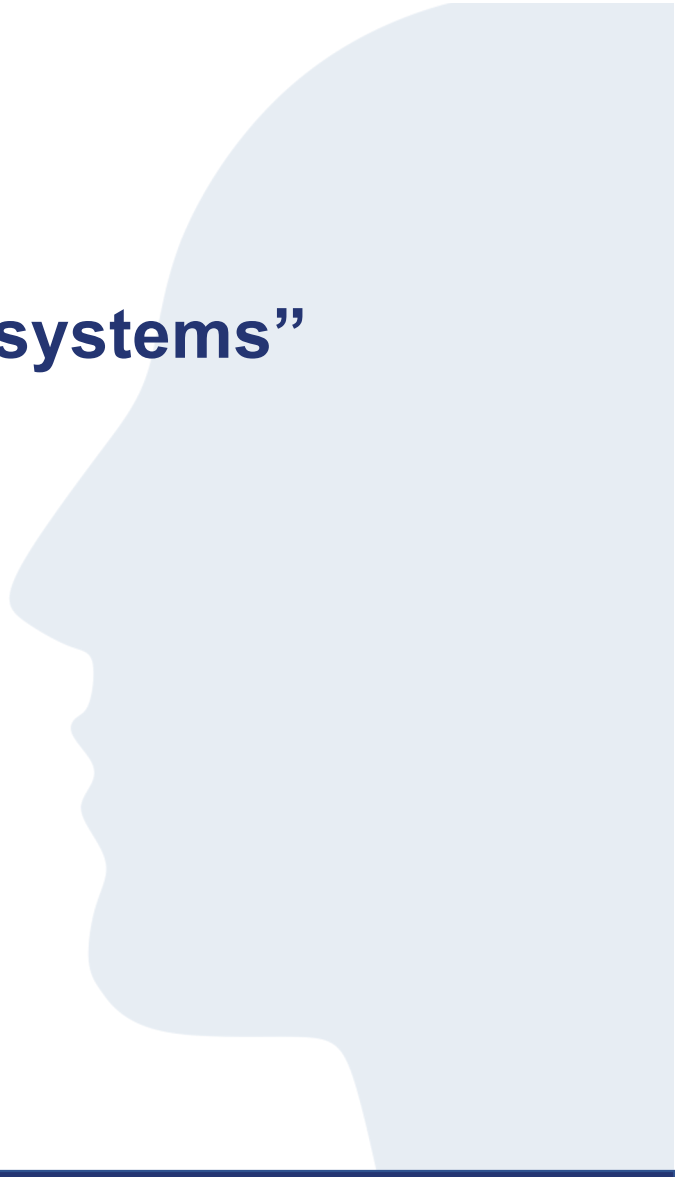
Recap

- **Last Week**: What is AI?
  - “The question of how an intelligent being is defined, is a long and difficult one.”
  - Rationality behavior is about doing the right thing but is it a very good model of reality?
  - We saw AIs doing great things but also how they can be tricked

Today

“How do we think about/define AI systems”

- What is an AI System?
- Environments
- Characteristics of Environments
- Agents
- The problem with Rational Agents
- Types of Agents



What is an AI System

AI System

An AI system can be defined as the study of (rational) agents and their environments.

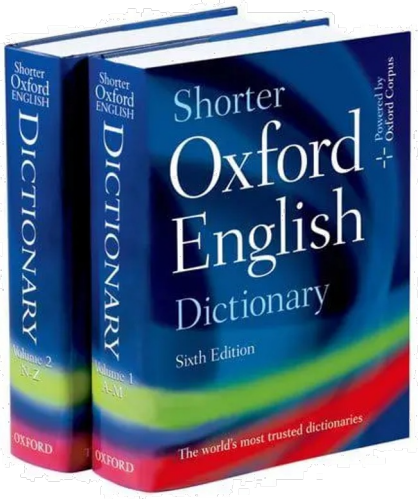
Has two parts:

1. **Environment**
2. **Agent**

What is an AI System
What is an Environment

AI System

An AI system can be defined as the study of (rational) agents and their environments.



1. **Environment**
   - Definition from Oxford Dictionary:
     „*environment is the surroundings or conditions in which a pers on, animal, or plant lives or operates*“.
   - In the context of AI, an environment is simply the surrounding of an (AI) agent and is where the agent operates.
   - An environment does not have to be real. It could also be artific ial.

Examples

- Selfdriving cars:
  Streets, traffic, weather, road signs, pedestrians, …
- Chess: The chess boars, the chess pieces
- …

Characteristics of Environments (I)

Dimensions to characterize different environments

Environment	Discrete?	Observable?	Static?	Single Agent?	Accessible?	Deterministic?	Episodic?
Chess							

Characteristics of Environments (I)

Dimensions to characterize different environments

Environment	Discrete?	Observable?	Static?	Single Agent?	Accessible?	Deterministic?	Episodic?
Chess							

Discrete vs. Continuous

Has the environment (a limited/countable) number of distinct, clearly defined states then it is called discrete; otherwise it is continuous.

A self-driving car is an example of a continuous environment

Observable vs. Partially Observable or Unobservable

Is it possible to determine the complete state of the environment at each time point, then it is called observable; otherwise it is only partially observable or even unobservable.

A self-driving car is an example of a partially observable environment, as the taxi cannot directly observe the intentions of pedestrians. Poker is another example of a partially observable environment.

Characteristics of Environments (II)

Dimensions to characterize different environments

Environment	Discrete?	Observable?	Static?	Single Agent?	Accessible?	Deterministic?	Episodic?
Chess	discrete	observable					

Static vs. Dynamic

If the environment does not change while an agent is acting, then it is static; otherwise it is dynamic.

Taxi driving is an example of a dynamic environment whereas Crossword puzzles are an example of a static environment.

Single Agent vs. Multiple Agents

The environment may contain other agents which may be of the same or different kind as that of the agent.

E.g. other taxi driver or poker players. The agent design problems in the multi-agent environment are different from single agent environment.

Characteristics of Environments (III)

Dimensions to characterize different environments

Environment	Discrete?	Observable?	Static?	Single Agent?	Accessible?	Deterministic?	Episodic?
Chess	discrete	observable	static	Multi-agent			

Accessible vs. Inaccessible

If an agent can obtain complete and accurate information about the state's environment, then such an environment is called an accessible environment else it is called inaccessible.

Mathematical theorem proving: In principle, all premises and inference rules are known; no hidden state

Deterministic vs. Non-deterministic/Stochastic

If the next state of the environment is completely determined by the current state and the actions of the agent, then the environment is deterministic; otherwise it is non-deterministic.

Chess: When you move a piece, the resulting board configuration is fully determined by the rules — no randomness involved.

Self-driving cars: Turning the steering wheel a certain angle may have slightly different effects depending on road friction, wind, etc. Often co-occurs with non-determinism: if you cannot fully observe the state, you also cannot perfectly predict outcomes

Difference between Accessible and Observable

We have an AI system, that...

- Should recognise objects in an image
- Can perceive the whole image
- Has no additional data or additional images

Is this environment observable? Is this environment accessible?

Accessibility concerns *the environment itself,* whether the information *exists* and can in principle be obtained.

Observability concerns *the agent's sensors,* whether they can actually *perceive* that information.

Characteristics of Environments (IV)

Dimensions to characterize different environments

Environment	Discrete?	Observable?	Static?	Single Agent?	Accessible?	Deterministic?	Episodic?
Chess	discrete	observable	static	Multi-agent	accessible	deterministic	

Episodic vs. Non-episodic/Sequential

In an episodic environment, each episode consists of the agent perceiving and then acting. The quality of its action depends just on the episode itself. In sequential environments the agent requires memory of past actions.

Why you need to know all that (it is less about the correct classification)

- Different environments need different agents
- Not every algorithm works in every environment
  - e.g. some algorithm do not work with uncertainty, i.e., need perfect information

Characteristics of Environments

How to specify/define Characteristics via Questions

- Do we have a finite number of states?
- Can we observe the game perfectly?
- Are there states/actions we cannot observe?
- Is the environment changing?
- Are actions irreversible?
- Does the environment contain other agents?
- Maybe some agents have other forms than our agent?
- Do we have an opponent?
- …

- Can we observe the complete environment?
- Is our observation accurate?
- Does actions always result in the same next state?
- Do we have randomness in our environment?
- Can we determine in which state we will end taking specific actions?
- Do we need memory of the past to determine the best action?

Characteristics of Environments
Easy and Difficult Environments

Szenario 1:

Static

- We do not pay attention to possible changes in the environment

Observable

- We can at least observe our initial state

Discrete

- Possible actions can be enumerated

Deterministic

- The expected outcome of an action is always identical to the true outcome

Szenario 2:

Dynamic

- The environment is changing all the time, even without acting

Partially Observable

- Do we know our current state

Continious

- So many possiblities

Stochastically

- We can make the same action multiple times in the same state but with different results

What is more realitstic? What is easier?

Characteristics of Environments

Try it out yourself. Imagine even other environments

Environment	Discrete?	Observable?	Static?	Single Agent?	Accessible?	Deterministic?	Episodic?
Chess	Discrete	Observable	Static	Multi-Agent	Accessible	Deterministic	Sequential
Solitaire							
Poker							
Self-Driving							
Medical Diagnosis							

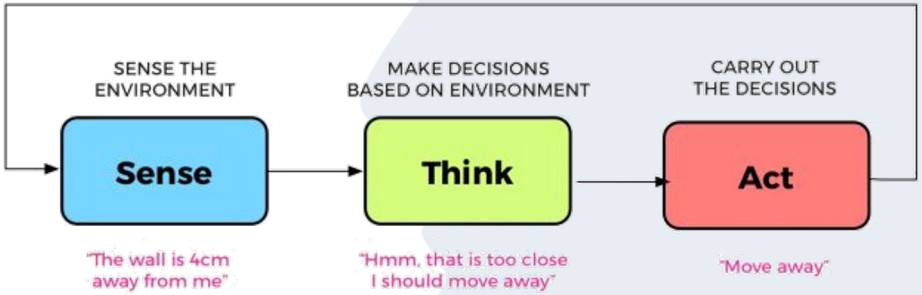
What is an AI Systems

What is an Agent

AI System

An AI system can be defined as the study of (rational) agents and their environments.

1. **Environment** ✓
2. **Agent**
   - perceives its environment (**Sense**)
   - makes decisions autonomous (**Think**)
   - acts upon the environments (**Act**)



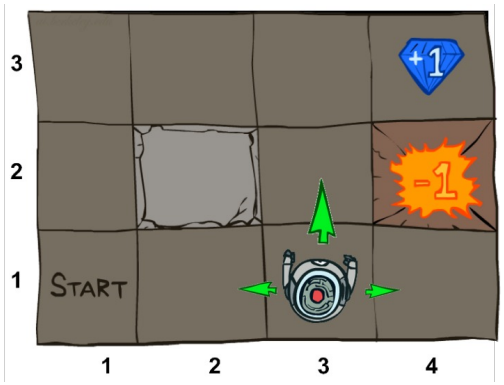
Rules of AI agents

1. An agent must be able to perceive the environment
2. The environments observations must be used to make decisions
3. The decisions should result in action
4. (The action taken by the agent must be rational)

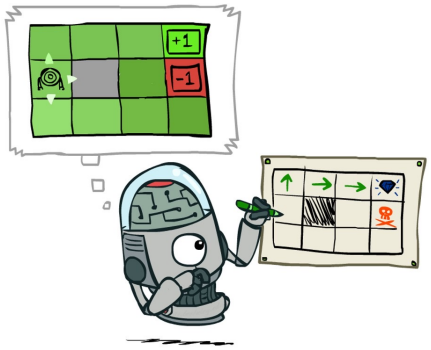
What is an AI Systems
What is an Agent

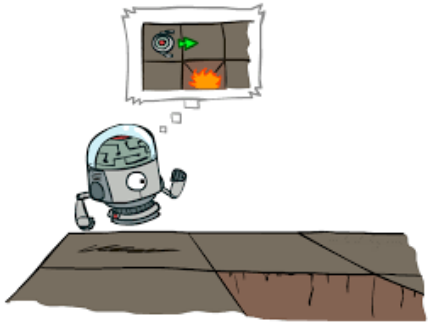
AI System

An AI system can be defined as the study of (rational) agents and their environments.



1. Perceive





2. Decide

3. Act

Images: Pieter Abbeel and Dan Klein, CS188 Intro to AI at UC Berkley

What is an AI Systems

What is an Agent

AI System

An AI system is an agent. An agent is anything that can be viewed as perceiving its environment through sensors and acting upon that environment through actuators. So:

AI = the study and design of (rational) agents that interact with their environments

Examples of agents

1. Humans:
   - Senses the world with their eyes, ears and other sensors
   - Thinks with their head
   - Acts with their hands, legs, vocal tract, and other actuators
2. Robots:
   - Senses the world with cameras, infrared and other sensors
   - Makes decisions based on inputs, rules or a program
   - Acts with motors for actuators
3. ...

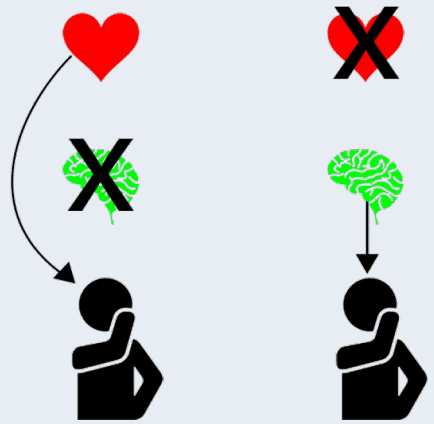
The Problem of Rationality

Recap: Rational actions

An action is rational, if it maximizes the performance and yield the best positive outcome for the agent.

A rational agent is an agent that does the **right thing**

What is the "right thing"?



Human behavior: Irrational behavior ignores thinking, while over rational ignores feelings

Be careful! Rational Agents are not omniscient

- An omniscient agent would know the actual outcome of its actions

Rational Agents are not perfect

- Rationality maximizes the expected performance
- This may not the optimal outcome
- e.g. playing in the lottery has a negative expected outcome, so it would be better to not play, but...

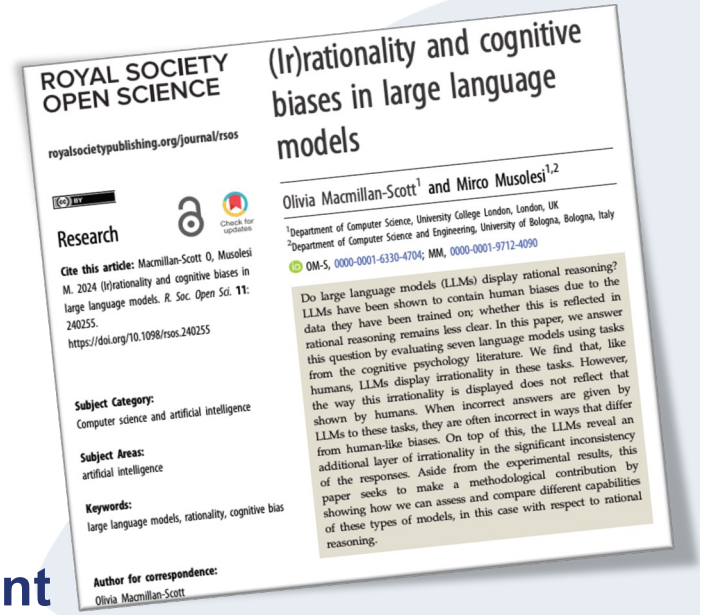
The Problem of Rationality

How to measure performance

- A funtion that evaluates a sequence of actions
- Is task-dependent
    - A vacuum cleaner has a different performance criteria than a self-driving car

General rule in designing a performance measurement

- Design the performance measure based on the desired outcome, not the desired agent behaviour


ROYAL SOCIETY OPEN SCIENCE

royalsocietypublishing.org/journal/rsos

Research

Cite this article: Macmillan-Scott O, Musolesi M. 2024 (Ir)rationality and cognitive biases in large language models. *R. Soc. Open Sci.* **11**: 240255. https://doi.org/10.1098/rsos.240255

Subject Category:
Computer science and artificial intelligence

Subject Areas:
artificial intelligence

Keywords:
large language models, rationality, cognitive bias

Author for correspondence:
Olivia Macmillan-Scott

(Ir)rationality and cognitive biases in large language models

Olivia Macmillan-Scott[1] and Mirco Musolesi[1,2]

[1]Department of Computer Science, University College London, London, UK
[2]Department of Computer Science and Engineering, University of Bologna, Bologna, Italy

OM-S, 0000-0001-6330-4704; MM, 0000-0001-9712-4090

Do large language models (LLMs) display rational reasoning? LLMs have been shown to contain human biases due to the data they have been trained on; whether this is reflected in rational reasoning remains less clear. In this paper, we answer this question by evaluating seven language models using tasks from the cognitive psychology literature. We find that, like humans, LLMs display irrationality in these tasks. However, the way this irrationality is displayed does not reflect that shown by humans. When incorrect answers are given by LLMs to these tasks, they are often incorrect in ways that differ from human-like biases. On top of this, the LLMs reveal an additional layer of irrationality in the significant inconsistency of the responses. Aside from the experimental results, this paper seeks to make a methodological contribution by showing how we can assess and compare different capabilities of these types of models, in this case with respect to rational reasoning.

Some possible performance criteria:

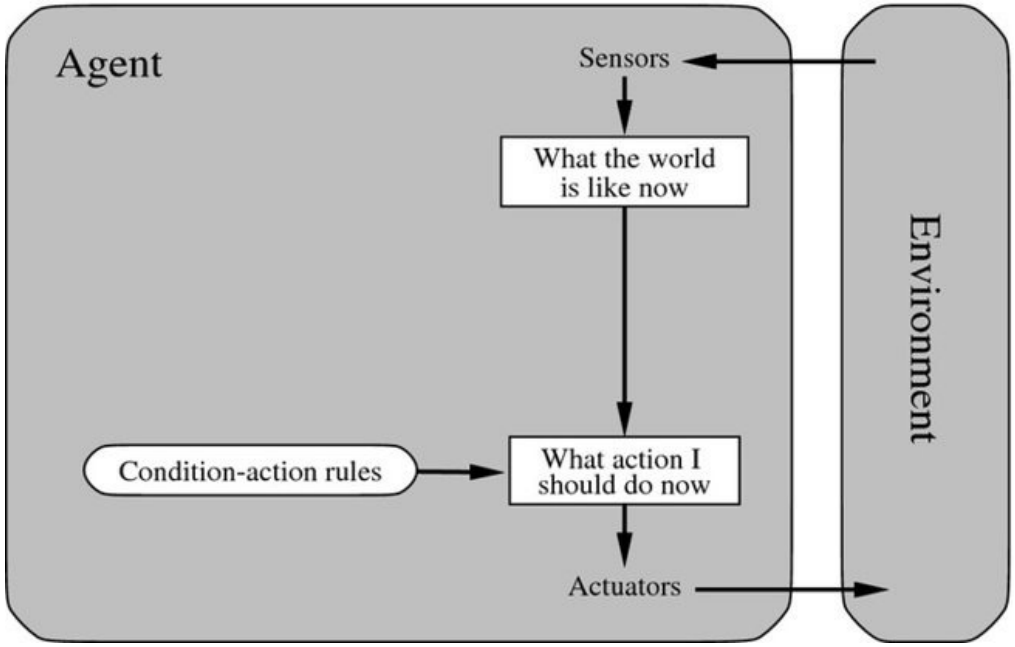
Vacuum cleaner: clean floors, energy efficency, cleaning time, noise,…

Selfdriving car: Safety, comfort, time efficency, routing,…

Types of Agents
Reflex Agent

Reflex Agent

Select action on the basis of only the current percept but ignores the percept history.



Thermostats may act according to a condition–action rule, such as:

If temperature < 20°C → turn on the heater.

If temperature ≥ 20°C → turn off the heater.

Types of Agents
Reflex Agent

Reflex Agent

Select action on the basis of only the current percept but ignores the percept history.

- Implemented through condition-action rules, i.e. "map state to action"
- Makes a very bad chess player
  - Does not look on the board, only at the last move played
  - No explicit goal of winning the game

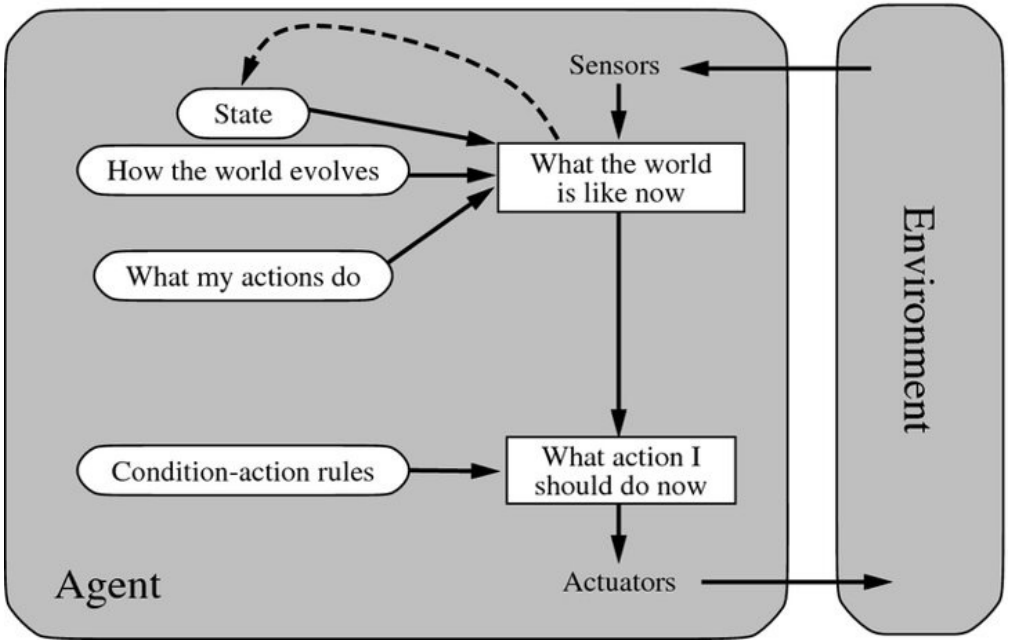
Problem

- Very limited in ist decision making
- No knowledge about anything which the agent cannot actively perceive
- Can become very hard to handle (store, update,…) in complex environments

Types of Agents
Model-based Agent

Model-based Agent

These agents choose their actions like reflex agents do, but they have a better comprehensive view of the environment, i.e. keep track of the world state



Warehouse robots keep track of the inventory positions and paths

Types of Agents
Model-based Agent

Model-based Agent

These agents choose their actions like reflex agents do, but they have a better comprehensive view of the environment, i.e. keep track of the world state

- Input is not only interpreted, but mapped into an internal state (and state dynamics) description (a world model)
- Makes a better chess agent
- could keep track of the current board situation when its percepts are only the moves

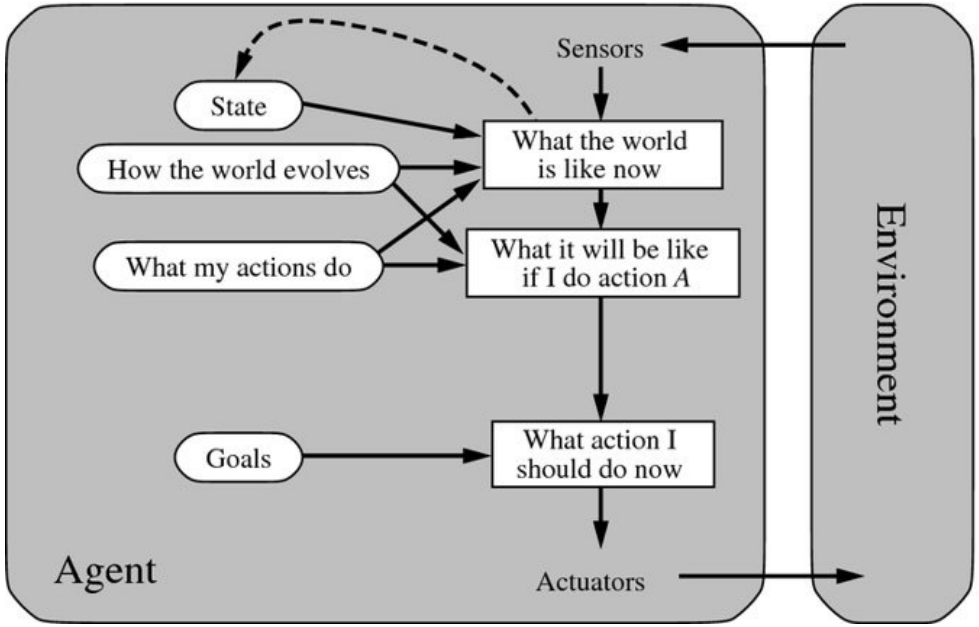
Problem

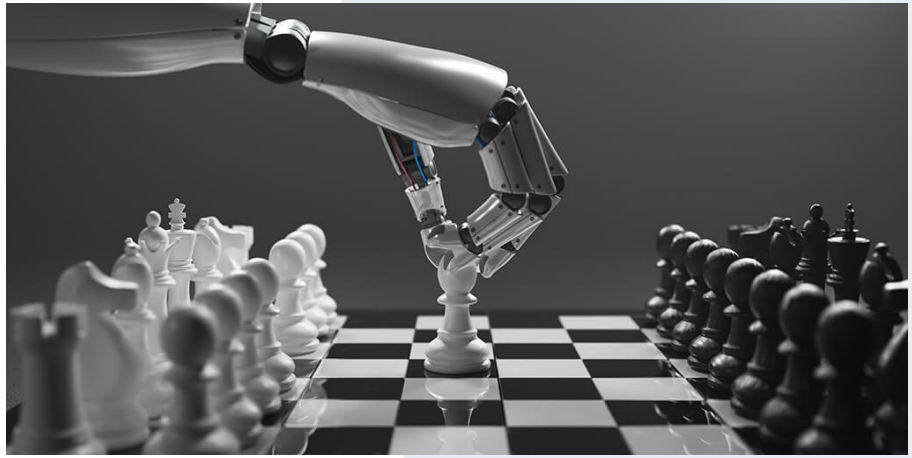
- How does my actions affect the world?
- What world model do I desire?

Types of Agents
Goal-based Agent

Goal-based Agent

These agents build on the information that a model-based agent stores but in addition knows what states are desirable..





The goal of a chess-playing agent is to chekmate the opponent

Types of Agents
Goal-based Agent

Goal-based Agent

These agents build on the information that a model-based agent stores but in addition knows what states are desirable (but not in a gradual way).

- The agent knows what states are desirable and will try to choose actions accordingly
- Main difference to previous approaches is that it takes decision-making into account
  - e.g. "What will happen if I do such-and-such?",
  - "What will make me happy?"

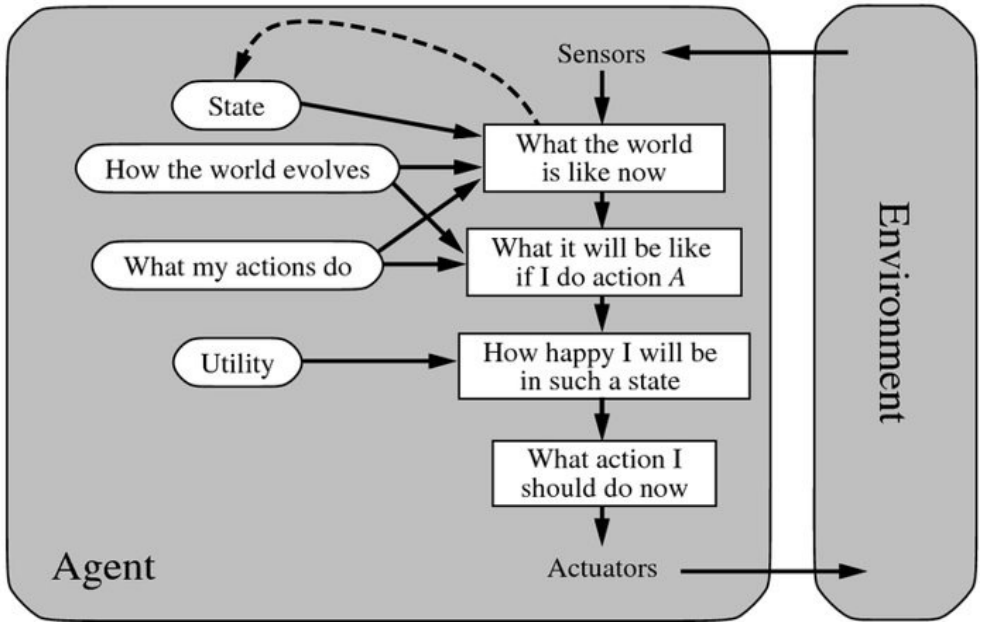
Problem

- Things become difficult when long sequences of actions are required to find/achieve a goal

Types of Agents
Utility-based Agent

Utility-based Agent

Instead of providing goals, utility-based agents use a utility function for providing a way to rate each action/scenario based on the desired result



With a smart home management AI system, we may want to maintain comfort while minimizing energy cost.

Types of Agents
Utility-based Agent

Utility-based Agent

Instead of providing goals, Utility-based agents use a utility function providing a way to rate each action/scenario based on the desired result

- Goals provide a binary distinction, while a utility function provides a contiuous scale
- Can help selecting between conflicting goals (e.g. is speed or safety more important)
- Certain goals can be reached in different ways
  - "Alle Wege führen nach Rom"
  - Some ways are quicker, safer, more reliable,… (have a higher utility)

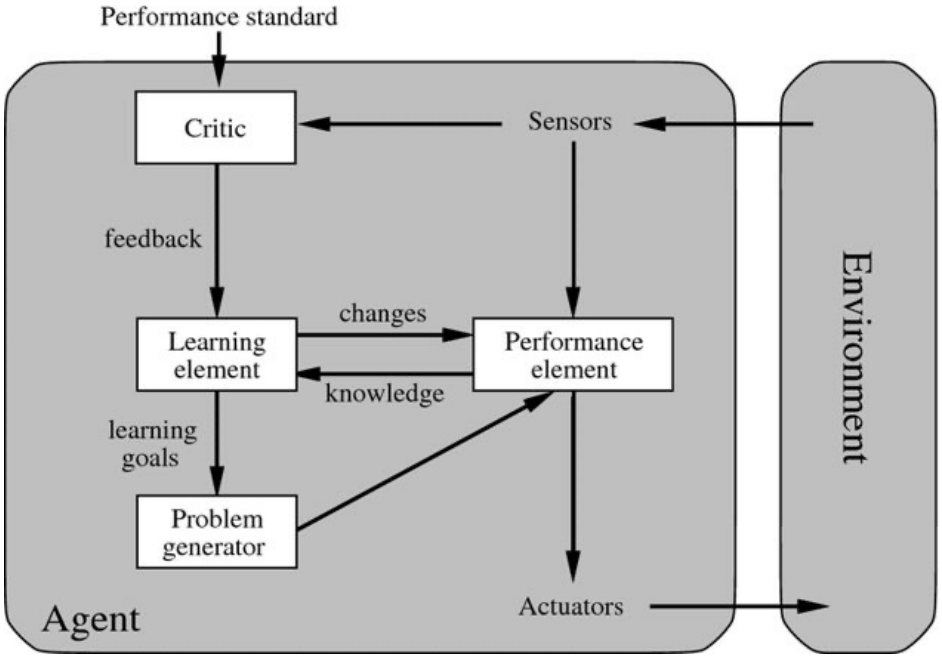
Utility function

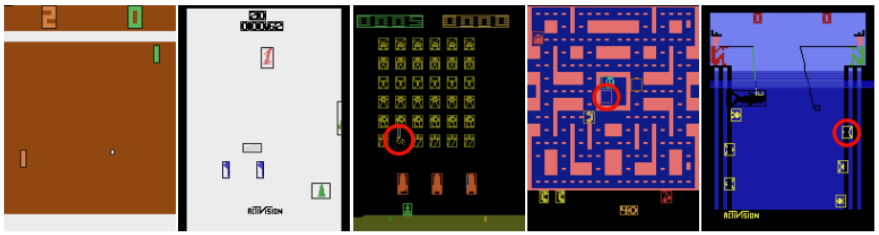
- Maps a state or sequence of state onto a real number

Types of Agents
Learning Agent

Learning Agent

These agents employ an additional learning element to gradually improve and become more knowledgeable over time about an environment.





Learning to play Atari games

Types of Agents
Learning Agent

- A learning agent can learn form its past experience,
  - i.e. is able to adapt automatically based on its experience
- Is more robust toward unknown environments
- A learning agent has four conceptual components

1. **Learning Element**:
   It is responsible for making improvements by learning from the environment
2. **Critic**:
   Gives feedback, describing how well the agent is doing with respect to a fixed mea surement
3. **Performance Element**:
   It is responsible for selecting actions
4. **Problem Generator**:
   Responsible for suggesting actions that will lead to new experiences

Reflex agent:
reacts.

Model-based agent:
remembers.

Goal-based agent:
plans.

Utility-based agent:
optimizes.

Learning agent:
improves itself over time.

How to Make Agents Intelligent

Intelligent agents make intelligent actions

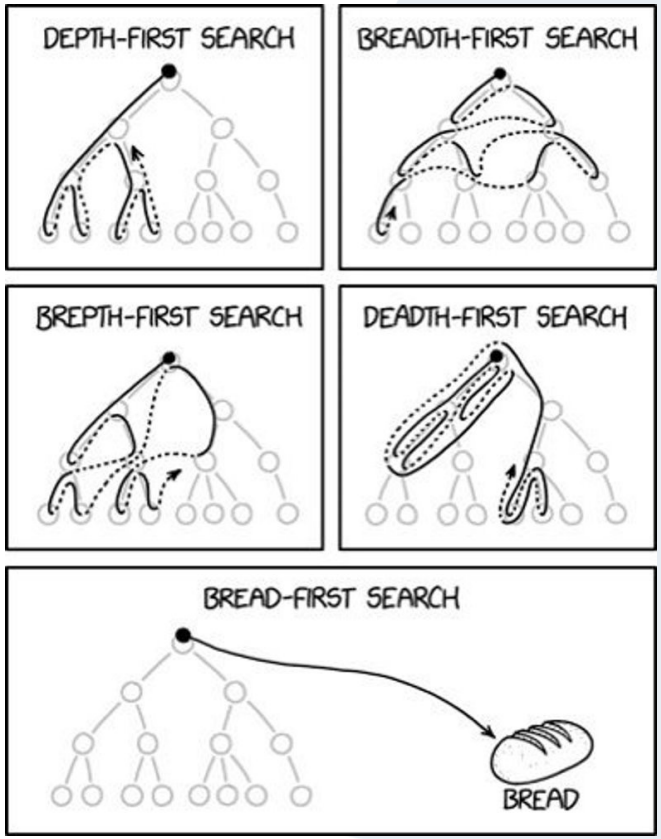
…but how do we decide what an intelligent action is and how do we select them?

Approaches

- Search Algorithms
- Reinforcement Learning
- Genetic Algorithms
- …

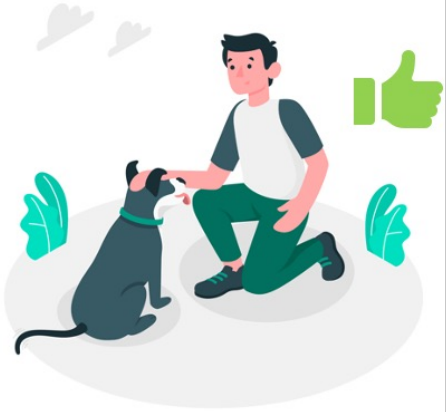
How to Make Agents Intelligent
Search Algorithms

- Understand/Define "finding a good action" as a search problem and use search algorithms
- Spoiler: Most common search algorithms are tree-based
    - Bread-First is none of them



Topic of the next lectures

How to Make Agents Intelligent
Reinforcement Learning



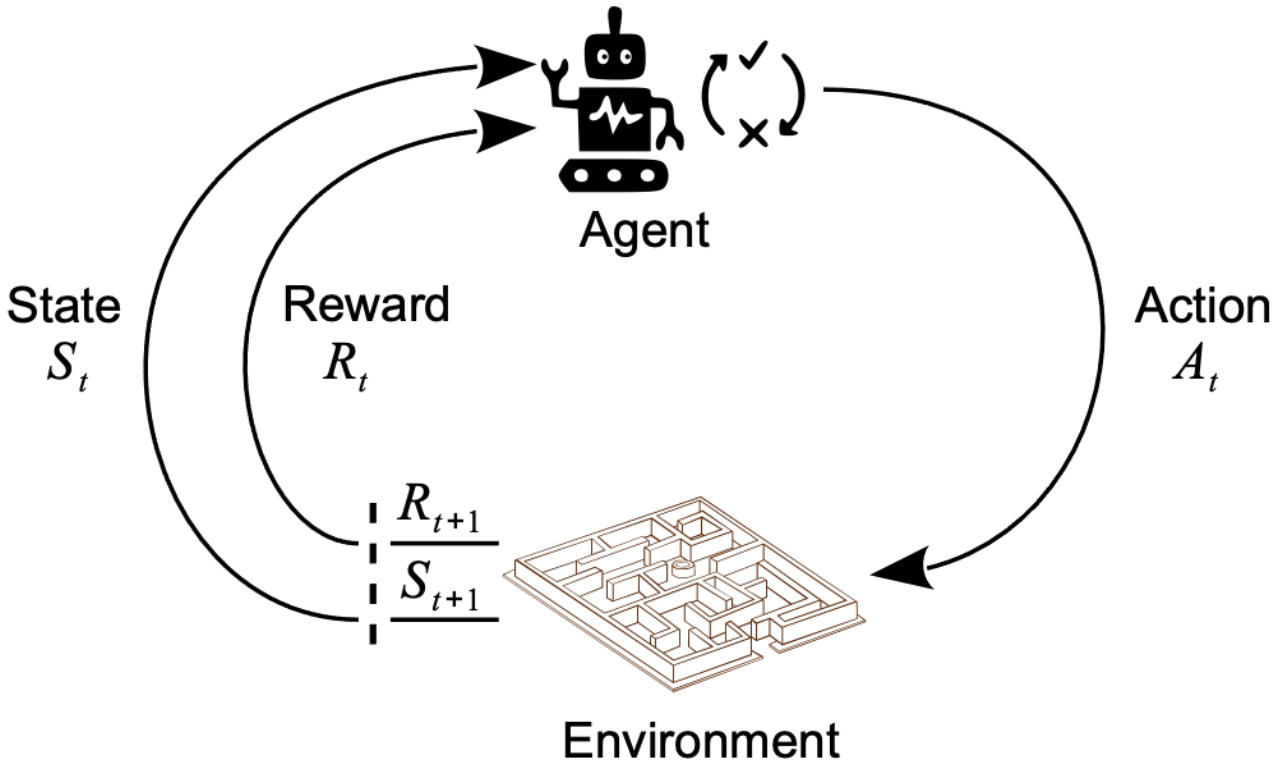
Conditioning of Animals
Designed by slidesgo (freepik)

Pain Signal when Touching a Hot Plate
Designed by macrovector and brgfx (freepik)

- Developed within the field of psychology
- Trial and Error
- Reactions/Actions are based on our observation and experience

How to Make Agents Intelligent

Reinforcement Learning Loop



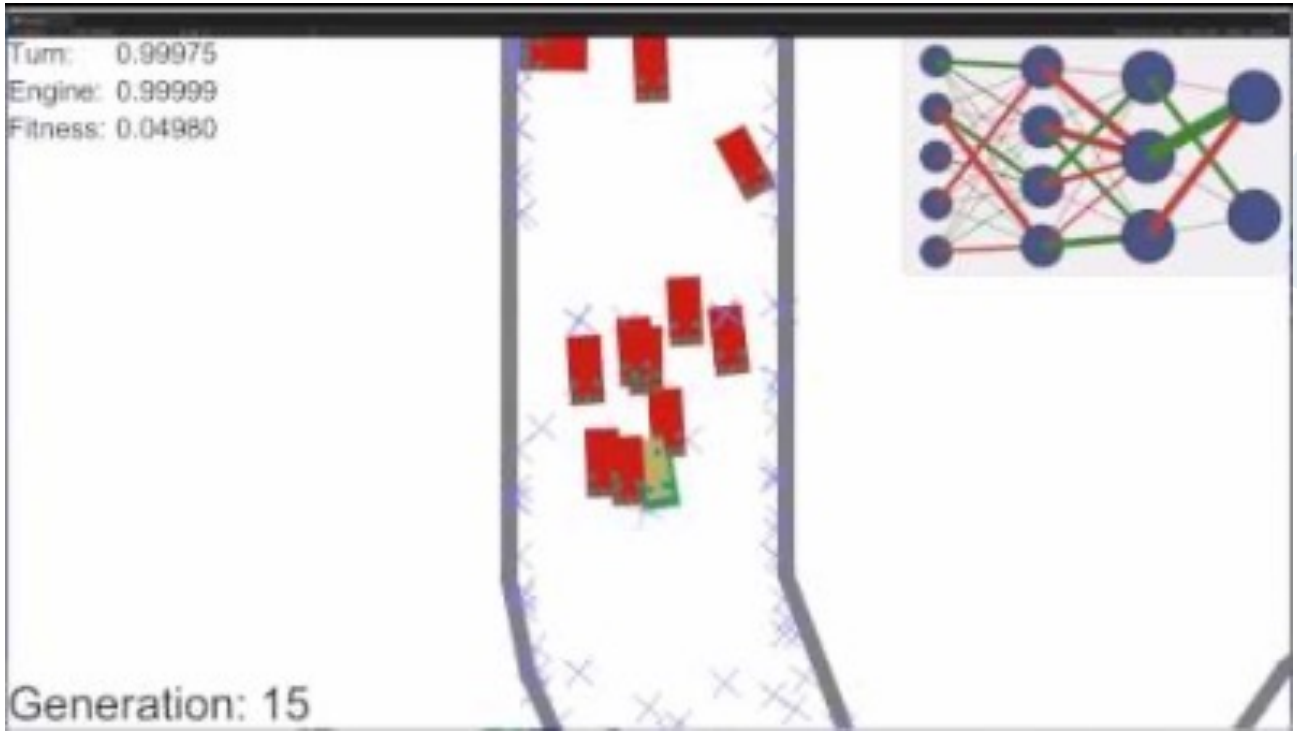
Interaction Loop between Agent & Environment

Part of the ML block (later this semester)

How to Make Agents Intelligent

Genetic Algorithms

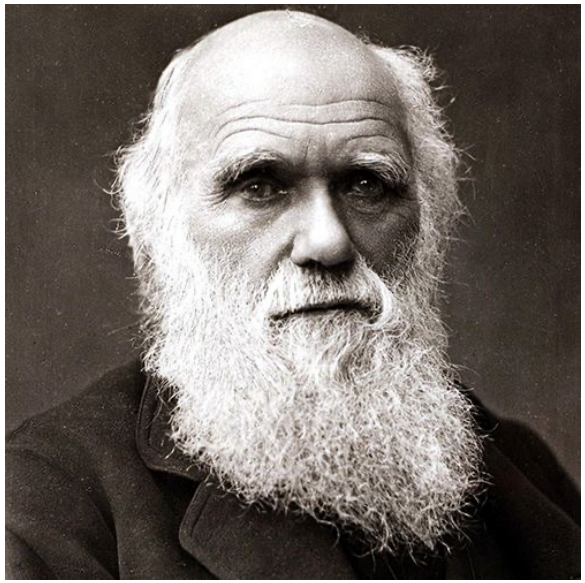
Survival of the fittest inspired by Charles Darwin's theory of natural evolution

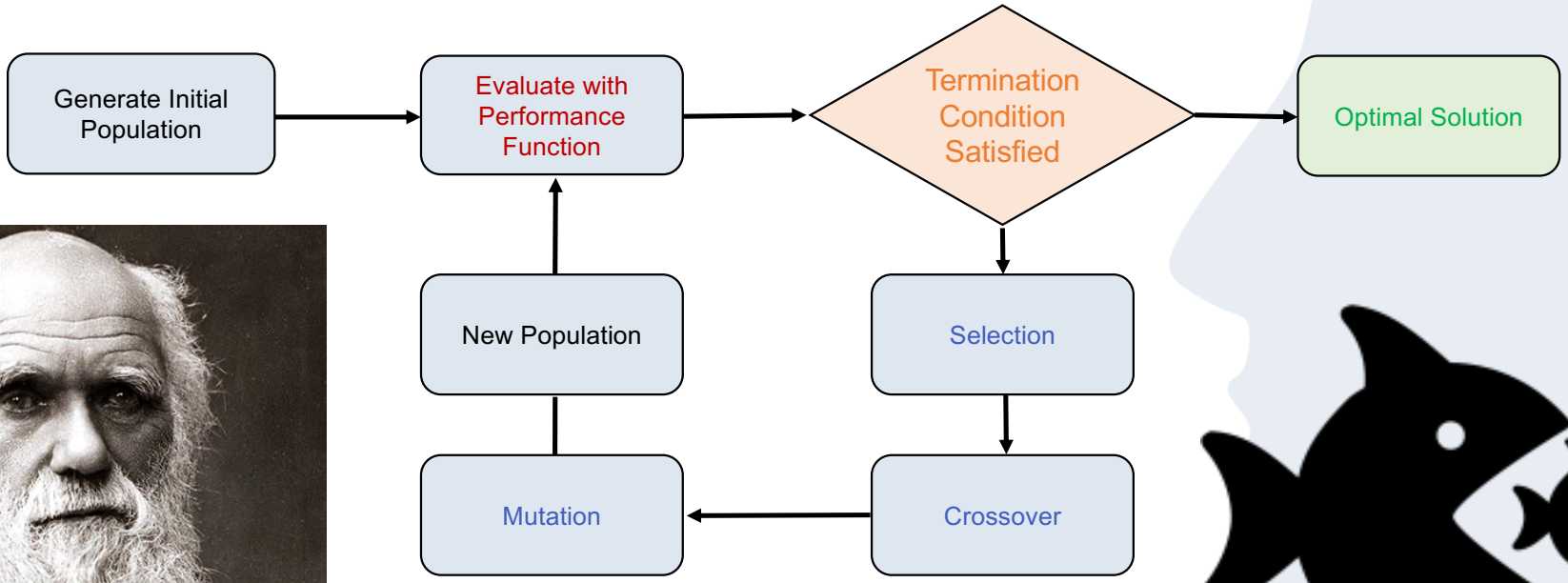


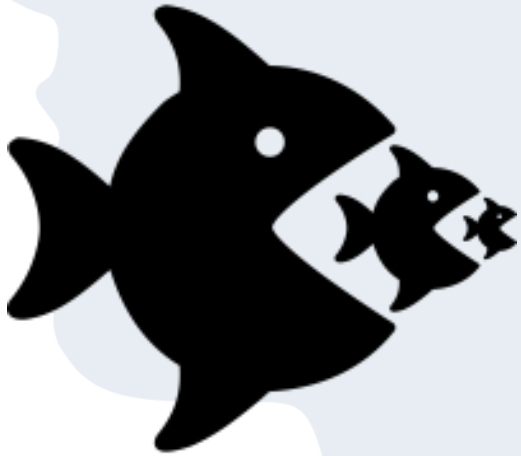
https://www.youtube.com/watch?v=Aut32pR5PQA

How to Make Agents Intelligent
Genetic Algorithms

Survival of the fittest inspired by Charles Darwin's theory of natural evolution







How to Make Agents Intelligent
Genetic Algorithms

Why is sexual reproduction so prevalent in nature when asexual reproduction seems far more efficient at promoting fitness?

A Livnat, C Papadimitriou, J Dushoff, MW Feldman, A mixability theory for the role of sex in evolution. *Proc Natl Acad Sci USA* **105**, 19803–19808 (2008)

A Livnat, C Papadimitriou, N Pippenger, MW Feldman, Sex, mixability, and modularity. *Proc Natl Acad Sci USA* **107**, 1452–1457 (2010)

Christos as featured in LogiComix. Image courtesy of Apostolos Doxiadis, Christos Papadimitriou, Alecos Papadatos, and Annie di Donna, 2009, Logicomix Print Ltd. and Bloomsbury Publishing Plc.

How to Make an Agent Intelligent
How to Improve the Performance

How do humans improve?

- We perform a task
- We remember the results
- We learn based on experience
- We go back to step 1

How do we improve agents?

- We perform a task
- We save the results of the attempt
- We adjust based on the saved data
- We go back to step 1

Is there a difference?

Training in Machine Learning (Lecture 9+10) is very similar to this idea

Summary

- Agents and environments are the main building blocks of AI software
- An agent is an entitiy that takes decisions based on what it perceiv es
  from its environment
- There are several types of agents based on their degree of perceived intelligence
- An ideal agents always tries to maximize its performance

You should be able to:

Explain the fundamental structure of an agent and the difference in agent types

Given a description, identify agents and environments as well as their properties

Next Week: Problem-solving, Uninformed Search, Informed Search