COIT20277 Introduction to Artificial Intelligence

Week 4 - Lecture

- Reinforcement Learning
- Responsible AI

BE WHAT YOU WANT TO BE
cqu.edu.au

Acknowledgement of Country

I respectfully acknowledge the Traditional Custodians of the land on which we live, work and learn. I pay my respects to the First Nations people and their Elders, past, present and future

BE WHAT YOU WANT TO BE
cqu.edu.au

Acknowledgment

The contents of this lecture have been adopted from the following references:

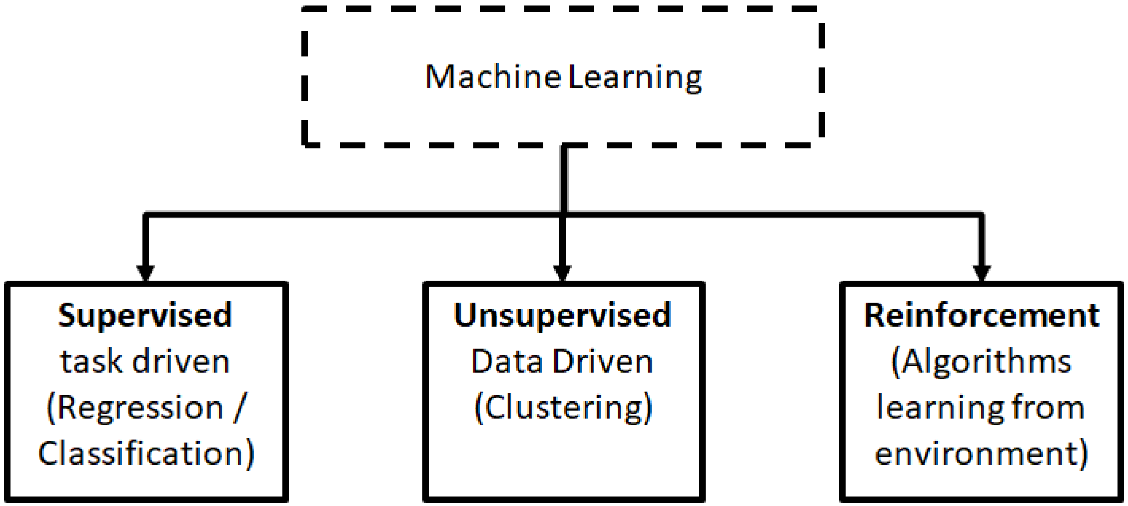
- Artificial Intelligence Programming with Python - From Zero to Hero, 2022, Perry Xiao, ISBN 978-1-119-82086-4:
  - Chapter 3.6
- Introduction to Responsible AI: Implement Ethical AI Using Python, Manure *et al.*, 2023, ISBN 978-1-4842-9981-4:
  - Chapter 1 and 2

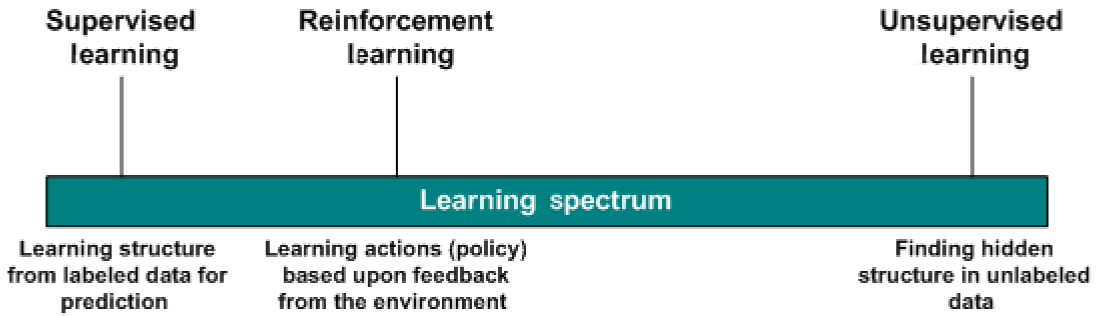
Topics

- **Part I: Reinforcement Learning**
  - What is Reinforcement Learning?
  - Key Terminology
  - Reinforcement Learning Algorithms
  - Applications of Reinforcement Learning
  - Popular Reinforcement Learning Platforms

- **Part II: Responsible AI**
  - Ethics in the Age of AI
  - Mitigating Bias and Discrimination
  - Privacy in the Age of Surveillance
  - Human-Centric Design

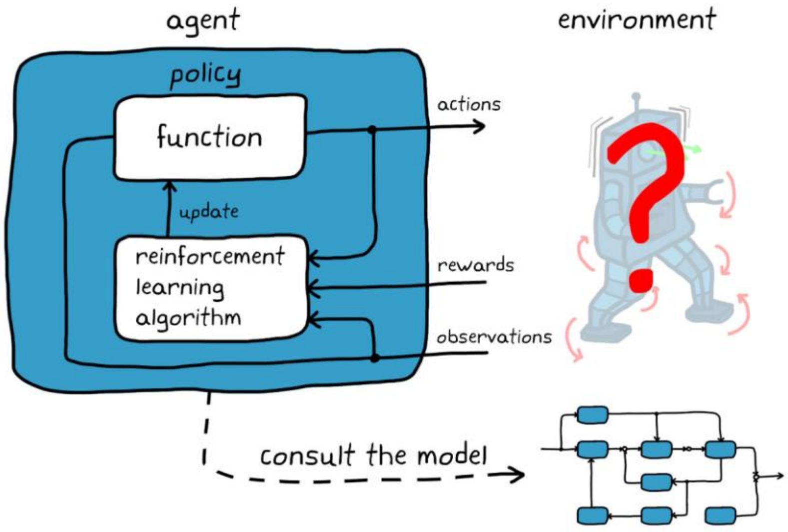
Context of Reinforcement Learning





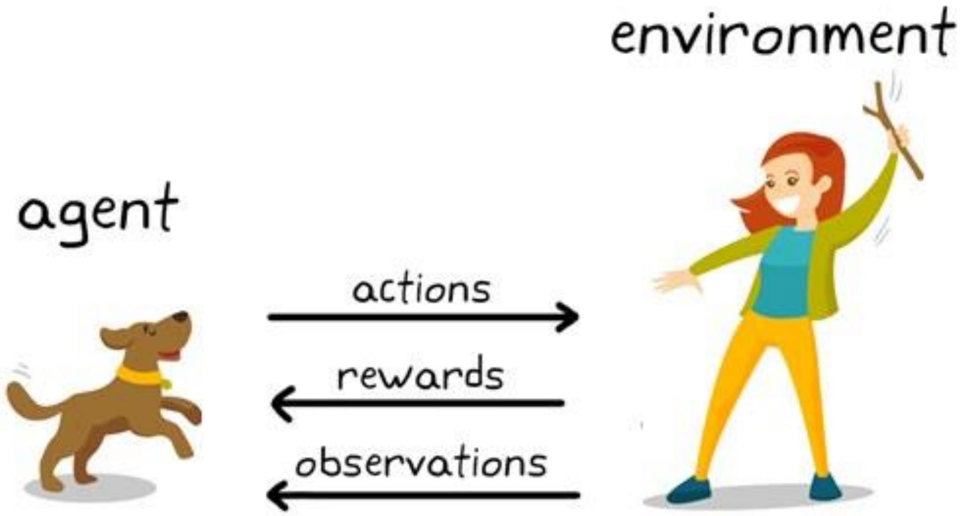
What is Reinforcement Learning?

- Reinforcement learning is a type of machine learning where an agent learns through trial and error in an interactive environment.
- The agent takes actions in the environment and receives rewards or punishments based on those actions.
- The agent's goal is to learn a policy that maps states to actions, so that it can maximize its long-term reward.



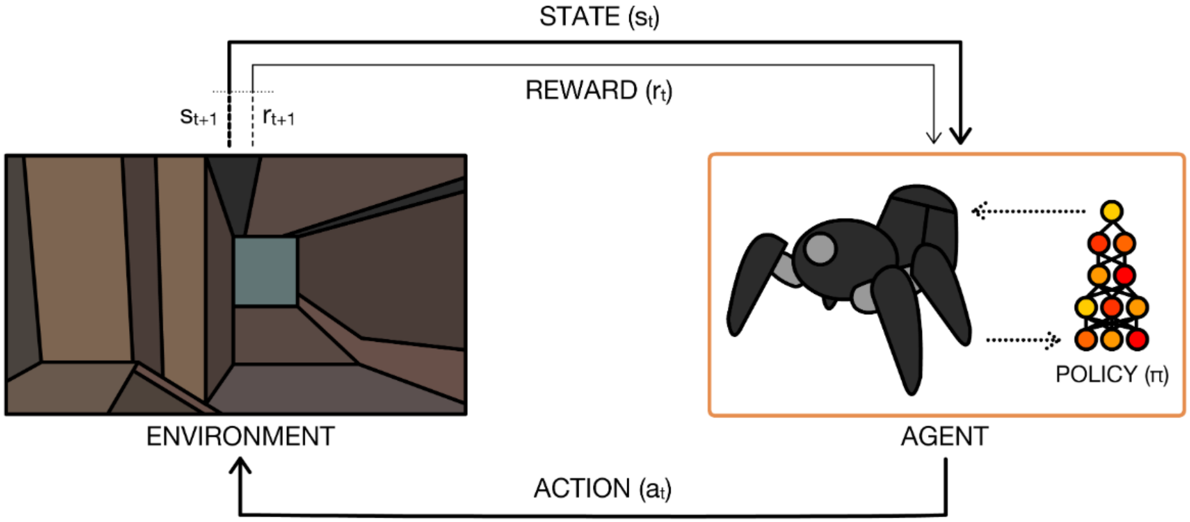
An Example

- Reinforcement learning is similar to how a dog learns tricks. The dog tries different actions, and every time it gets a right action, it gets a reward, or a treat.
- The next time, the dog learns to do the same thing again and gets the treat again. This is how reinforcement learning works.



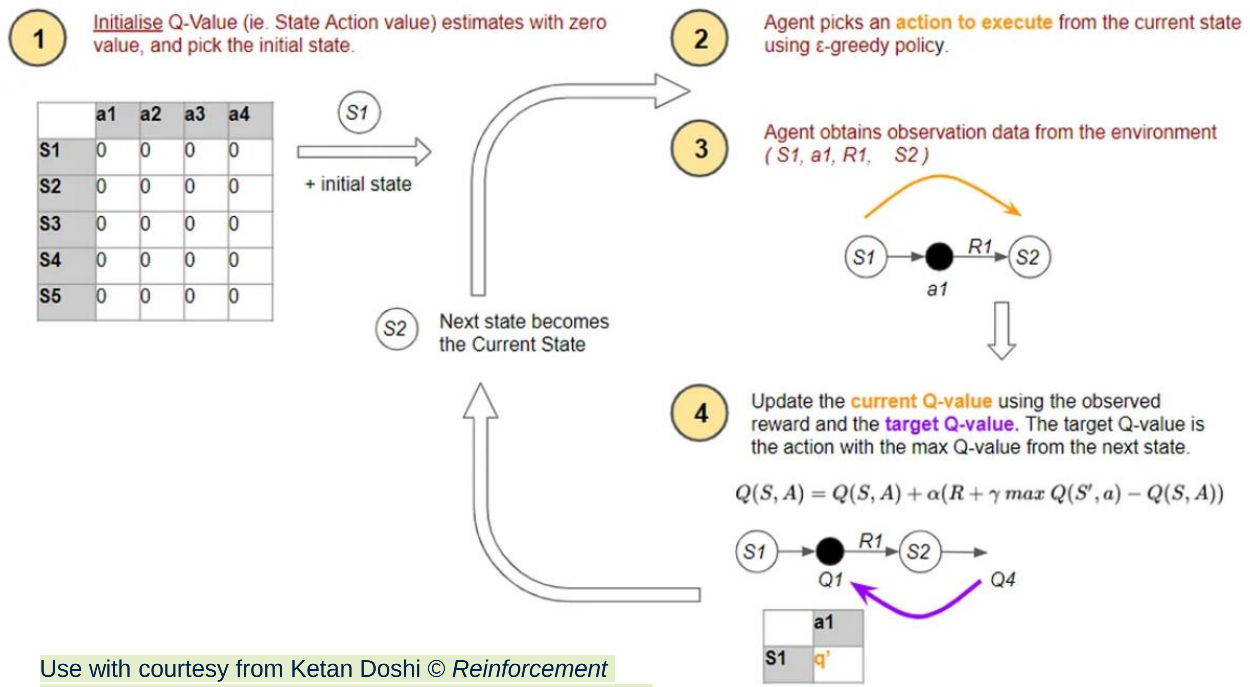
Key Terminology

- ***Environment:*** The physical world in which the agent operates.
- ***State:*** The current situation of the agent.
- ***Reward:*** Positive or negative feedback from the environment.
- ***Policy:*** The rules that change agent's state to actions.
- ***Value:*** Future reward that an agent would receive.



Reinforcement Learning Algorithms

- Q-Learning and SARSA are two commonly used model-free reinforcement learning algorithms.
- Q-learning is a ***model-free*** reinforcement learning algorithm.
- It does not require a complete model of the environment.
- It learns by interacting with the environment and receiving rewards.
- The goal of Q-learning is to learn a Q-value for each state-action pair.



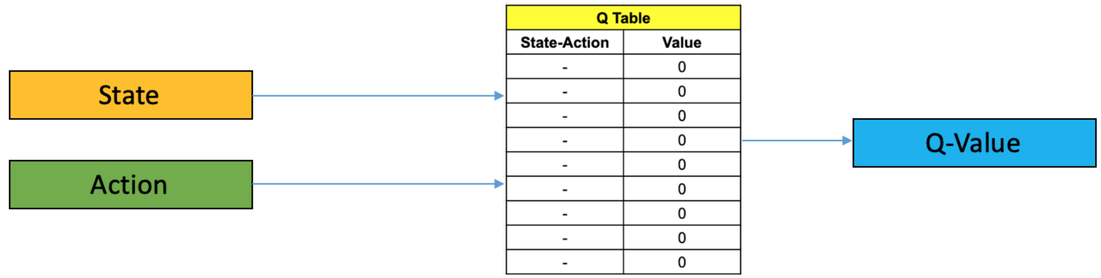
Use with courtesy from Ketan Doshi © *Reinforcement Learning Explained Visually (Part 4): Q Learning, step-by-step* Published in Towards Data Science Nov 28, 2020

Reinforcement Learning Algorithms (cont…)

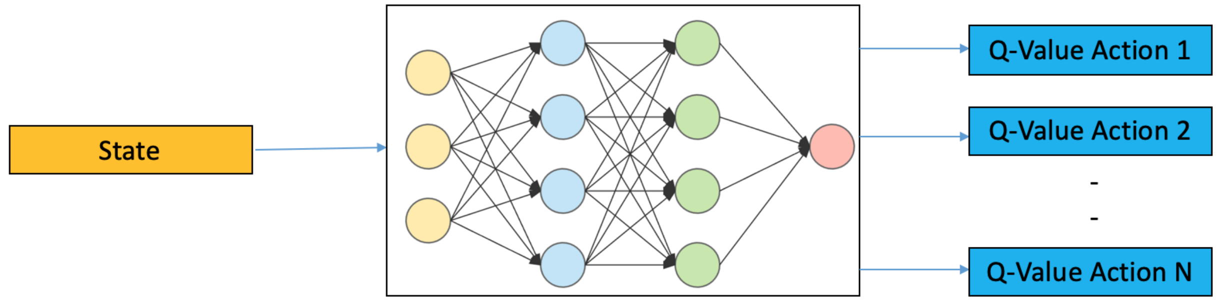
- Like Q-Learning, ***SARSA*** is *model-free and* learns directly from interaction, not a model of the environment.
- SARSA estimates the value of taking different actions in different states.
- It follows an "on-policy" approach, meaning it learns and improves the policy it's currently following.
- At each time step, SARSA selects an action based on its current policy (e.g., epsilon-greedy) and observes the reward and next state.
- SARSA updates its value estimates using the observed reward and the next state-action pair.
- It updates its policy based on the updated value estimates.
- SARSA is particularly suitable for online learning and environments where the policy needs to be continuously adjusted based on new experiences.

Reinforcement Learning Algorithms (cont…)

- More advanced algorithms such as ***Deep Q-Networks (DQN)*** and ***Deep Deterministic Policy Gradient (DDPG)*** can handle <u>unseen states and high-dimensional action spaces</u>.



Q Learning



Deep Q Learning

Use with courtesy from Ankit Choudhary © *A Hands-On Introduction to Deep Q-Learning using OpenAI Gym in Python*
Published in Analytics Vidhya on 21 August 2023

Q-Learning Example: Routing Problem

- Consider a simple routing problem with seven states (0-6).
- The goal is to find the best route from the start state (0) to the goal state (6).
- We can represent the problem as a graph, where nodes are states and edges are actions.
- Based on Figure 3.19, one can construct a corresponding matrix R, which indicates the reward values from a state to take an action to the next state, as shown in Figure 3.20 (next slide).
- The reward matrix R indicates the rewards for taking an action from one state to another.

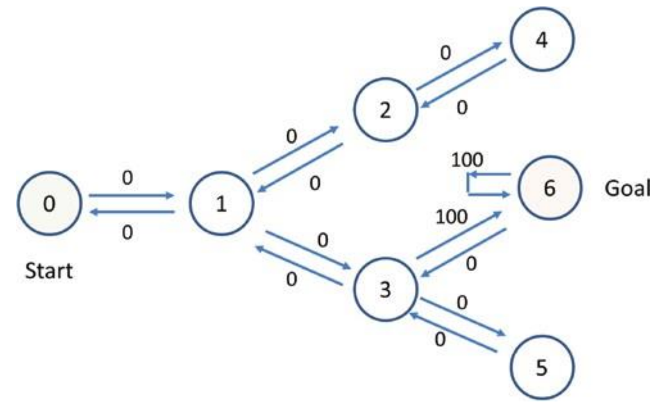


Figure 3.19: A simple routing problem with seven states, with 0 as the start state and 6 as the goal state (Xiao, P., 2022)

Q-Learning Example: Routing Problem (cont…)

- The value 0 means it is possible to go from one state to another state. The value -1 means it is not possible.
- The value 100 indicates reaching the Goal state; there are only two possibilities, from state 3 to state 6, and from state 6 to state 6.
- Based on R, one can also construct a similar matrix Q, and update the values of Q iteratively by using the following formula:

$$Q(s,a) = R(s,a) + \gamma \times max\big(Q(ns,aa)\big)$$

- where

  $Q(s,a)$=Q matrix value at state (s) and action (a).

  $R(s,a)$=R matrix value at state (s) and action (a).

  γ=the learning rate.

  $Q(ns,aa)$=Q matrix value at next state (ns) and all actions (aa).

  Max(.)=is the function to get the maximum values.

Action

$$R = \begin{array}{c} \\ 0 \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \end{array} \begin{array}{c} \begin{array}{ccccccc} 0 & 1 & 2 & 3 & 4 & 5 & 6 \end{array} \\ \left[\begin{array}{ccccccc} -1 & 0 & -1 & -1 & -1 & -1 & -1 \\ 0 & -1 & 0 & 0 & -1 & -1 & -1 \\ -1 & -1 & -1 & -1 & 0 & -1 & -1 \\ -1 & 0 & -1 & -1 & -1 & 0 & 100 \\ -1 & -1 & 0 & -1 & -1 & -1 & -1 \\ -1 & -1 & -1 & 0 & -1 & -1 & -1 \\ -1 & -1 & -1 & 0 & -1 & -1 & 100 \end{array}\right] \end{array}$$

State

Figure 3.20: The corresponding reward value R matrix of the routing problem (Xiao, P., 2022)

Topics

- **Importance of Responsible AI**
  - Ethics in the Age of AI
  - Mitigating Bias and Discrimination
  - Privacy in the Age of Surveillance
  - Human-Centric Design

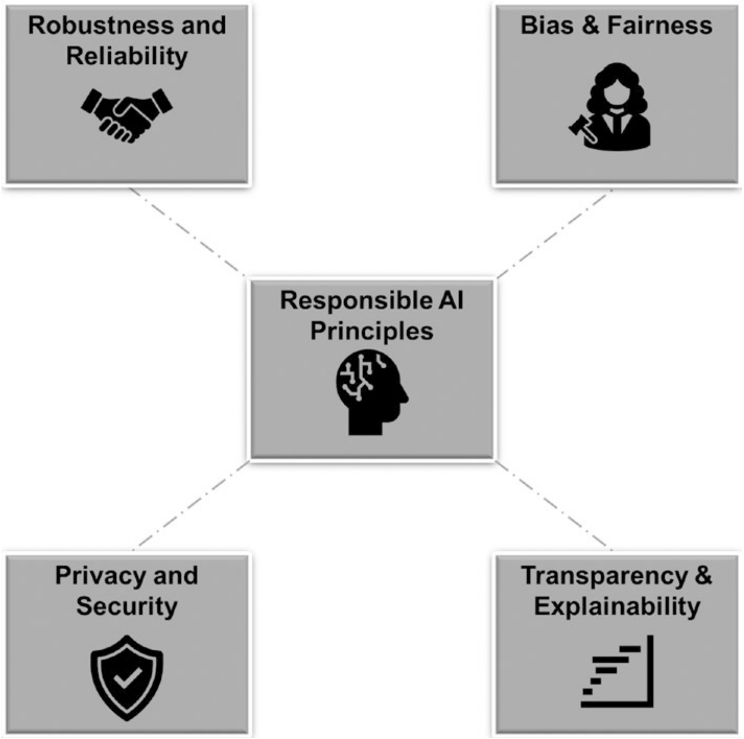


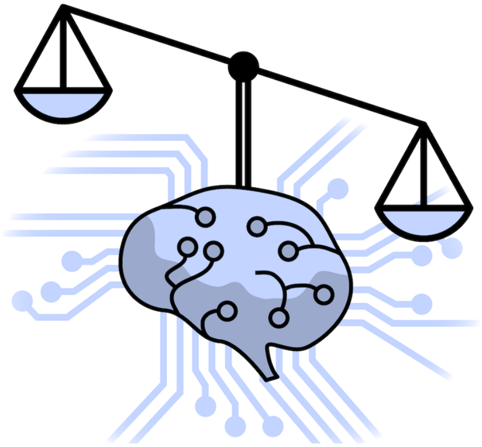
Figure 1-1. Evolution of artificial intelligence (Manure et al., 2023)

Ethics in the Age of AI

- AI has the potential to influence human lives, societies, and economies.
- AI also poses an ethical dilemma: the power to create and wield machines capable of decision making, learning, and autonomy.
- ***Responsible AI*** guides the development, deployment, and governance of AI technologies.
- It aligns technological innovation with societal values.
- It upholds ethical principles, accountability, and transparency throughout the AI lifecycle.
- It safeguards human well being and ensures equitable benefits for all.

Mitigating Bias and Discrimination

- A concern in the AI landscape is the potential for bias and discrimination in algorithms
- AI systems trained on biased data can perpetuate societal prejudices and exacerbate existing inequalities
- ***Responsible AI*** addresses this issue head-on, demanding rigorous data preprocessing, algorithmic transparency, and the pursuit of fairness
- It creates systems that reflect the diverse fabric of human society
- It bridges digital divides, ensuring that AI's impact is not marred by discriminatory practices
- It champions fairness and equity, paving the way for a future where technology is a tool of empowerment, rather than an agent of division.

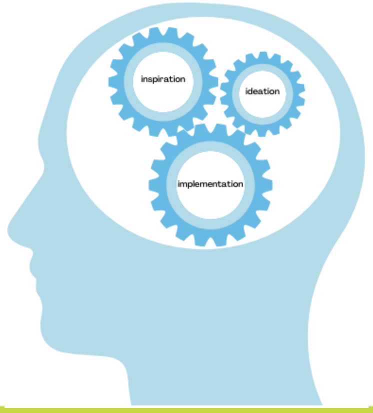


Privacy in the Age of Surveillance

- The era of digital advancement has resulted in an unparalleled rise in data creation, raising worries about individual privacy and the security of data.
- The insatiable appetite of AI for data necessitates a careful equilibrium between creativity and the protection of individual rights in its learning algorithms.
- ***Responsible AI*** highlights the significance of safeguarding data by promoting strong encryption, secure storage, and rigorous access management.
- It cultivates a sense of trust between technology and individuals.
- It empowers individuals to retain agency over their personal information while enabling organizations to harness data insights for positive transformations.
- It fortifies the pillars of privacy, ensuring that technological advancement does not come at the cost of individual autonomy.

Human-Centric Design

- Amidst the AI revolution, the concern that machines will replace human roles resonates strongly.
- ***Responsible AI*** dispels this notion by embracing a human-centric approach to technology.
- It envisions AI as an enabler, amplifying human capabilities, enhancing decision making, and fostering innovative synergies between man and machine.
- The importance of maintaining human oversight in AI systems cannot be overstated.
- It encourages the development of "explainable AI," wherein the decision-making processes of algorithms are comprehensible and traceable.
- It engenders trust and empowers individuals to make informed choices, thereby ensuring that AI operates in harmony with human values and goals.

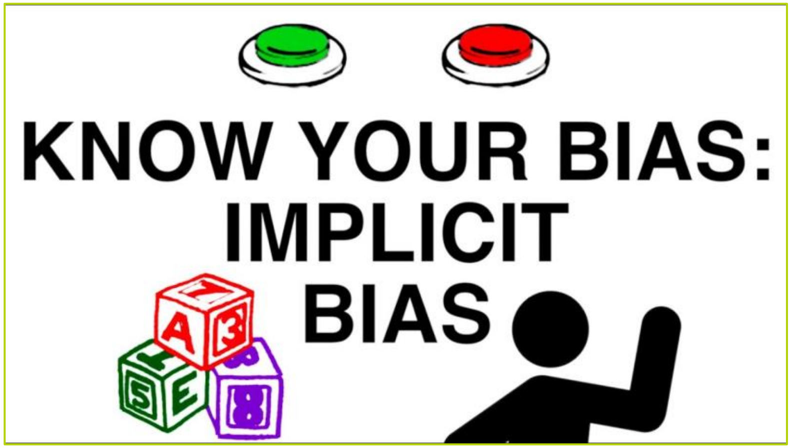


Bias in the AI Landscape

- ***Bias*** is the presence of systematic deviations that lead to inaccuracies or unfairness in decision making.
- Bias can impact various domains, from individual choices to complex models.
- Bias can stem from various sources, such as historical inequalities, flawed data-collection methods, or biased algorithms.
- Bias can distort outcomes and fairness, affecting individuals, groups, and societies.
- Bias can be detected and mitigated by technology, nurturing transparent and responsible AI.

Understanding Bias in Data and Models

- Bias in data and models emerges when data-collection or model-construction processes inadvertently favor certain groups, attributes, or perspectives over others.
- Bias in data and models can manifest in various ways, such as ***sampling bias***, ***measurement bias***, ***label bias***, ***algorithmic bias***, or ***outcome bias***.
- Bias in data and models can affect the performance, reliability, and validity of AI systems.
- Bias in data and models can be identified by analyzing the data distribution, the model assumptions, and the model outcomes.
- It can be addressed by implementing strategies that ensure equitable and unbiased decision making in artificial intelligence systems.



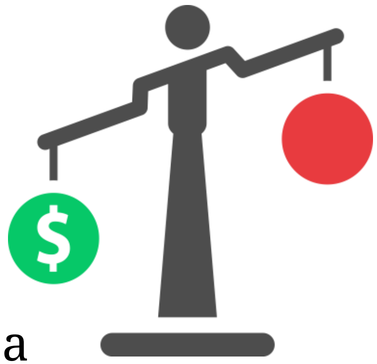
Recognising Bias for Creating Fair and Equitable Systems

- Recognising bias is the first step toward creating fair and equitable systems.
- It involves understanding the context, the stakeholders, and the objectives of the system.
- It requires defining and measuring fairness, which can be challenging and context-dependent.
- It entails evaluating the potential harms and benefits of the system for different groups and individuals.
- It enables the development of explainable AI, wherein the decision-making processes of algorithms are comprehensible and traceable.
- It fosters trust and accountability, ensuring that AI operates in harmony with human values and goals.

Techniques to Detect, Assess, and Mitigate Bias

- Detecting bias involves examining the data and the model for signs of systematic deviations or discrepancies.
- Detecting bias can be done using various methods, such as data visualization, statistical tests, or model evaluation metrics.
- Assessing bias involves quantifying the degree and the impact of bias on the system outcomes and fairness.
- It can be done using various measures, such as disparity, discrimination, or fairness metrics.
- Mitigating bias involves applying interventions to reduce or eliminate bias in the data or the model.
- It can be done using various techniques, such as data preprocessing, algorithm modification, or post-processing correction.

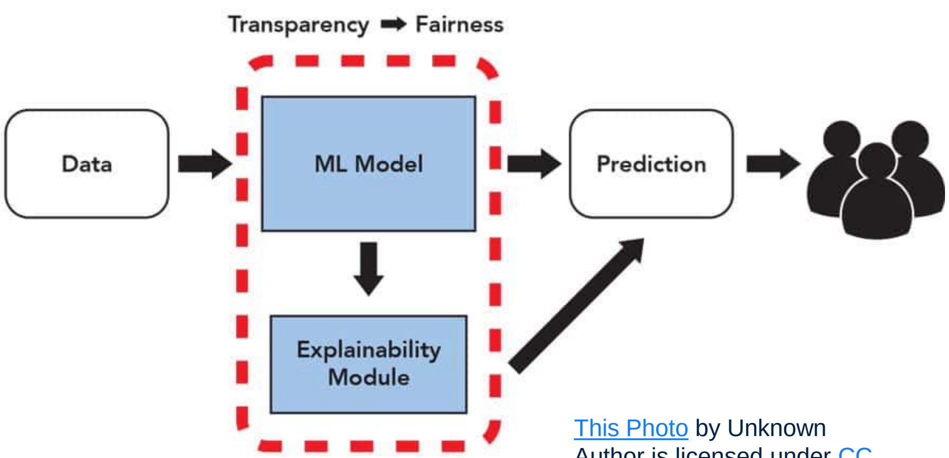
Trade-offs Between Model Complexity and Interpretability



- ***Model complexity*** refers to the number of parameters, features, or interactions that a model uses to learn from data.
- It can affect the accuracy, generalization, and efficiency of a model.
- ***Model interpretability*** refers to the ability to understand how a model makes predictions or decisions.
- It can affect the transparency, explainability, and trustworthiness of a model.
- There is often a ***trade-off*** between model complexity and interpretability, meaning that more complex models tend to be less interpretable, and vice versa.
- The trade-off between model complexity and interpretability can be balanced by using various methods, such as feature selection, regularization, or model-agnostic explanations.

Summary

- Bias and fairness are important concepts in the AI landscape.
- Bias and fairness can impact decision making across various domains, from individual judgments to automated systems.
- Bias and fairness can stem from various sources, such as historical inequalities, flawed data-collection methods, or biased algorithms.
- Bias and fairness can be detected, assessed, and mitigated by technology, nurturing transparent and responsible AI.
- Bias and fairness align with ethics, sculpting AI that champions diversity and societal progress.
- Bias and fairness require a balance between model complexity and interpretability, ensuring that AI systems are accurate, reliable, and understandable.



THANK YOU

TIME FOR DISCUSSION & QUESTIONS