4 Ethics in AI

Learning Objectives:

- Understand and appreciate ethical concerns around AI.
- Critically think about the cost and benefits of AI technology.

Background and Context

"In this era of profound digital transformation, it's important to remember that business, as well as government, has a role to play in creating shared prosperity—not just prosperity. After all, the same technologies that can be used to concentrate wealth and power can also be used to distribute it more widely and empower more people."

—Erik Brynjolfsson, Director of the MIT Initiative on the Digital Economy

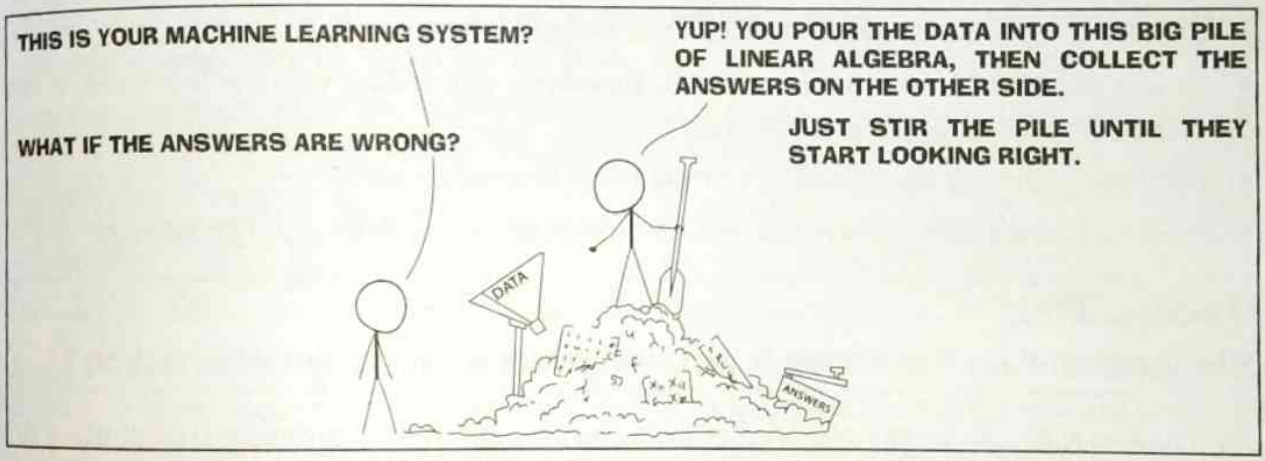


Fig. 4.1: Understanding Ethics (Source: *Machine Learning, XKCD*)

Activity 1

Watch the video: AI for Good

Scan QR code or visit:
https://www.youtube.com/watch?v=vgUWKXVv09Q

What have you understood from the video?
What are your learnings from it?

Activity 2 **Balloon Debate**

Divide yourselves into groups of four. Two groups will be given a theme related to advantages and disadvantages of various AI applications in different industries you researched about. One team will speak in favour of AI applications while the other will speak against it. The debate will go theme by theme and each member of both the teams will get a minute to speak. The chance to speak will alternate between the favouring and rebuttal team. Finally, only one team will remain in the balloon depending upon how convincing their points are. If any speaker speaks for more than a minute, their team will get disqualified. Each member will get 15 minutes to prepare. And your time starts now!

Imagine there are two families of four people out for a ride in a hot air balloon. Suddenly, the balloon starts moving towards the earth instead of staying airborne. To stabilize it, one family needs to take the parachute and go out of the balloon or else, it will come crashing down.

Who should be thrown out of the hot air balloon?

Reflect and Discuss:

- With the rise in AI applications replacing human workforce, do you consider it ethical to incorporate the use of AI in various jobs?
- How do you think income would be shared if AI is used in place of human workforce?
- AI will probably bring with it many health benefits. How will these health benefits be made accessible and available to all the people in the society?
- AI is a powerful tool in various fields. However, depending on how it is used, it can either be a boon or a bane. Discuss.
- How can learning opportunities for AI be extended to all?
- How will human beings ensure that they stay ahead of Artificial Intelligence?

Useful Tip:

"The important thing to remember is the consequences of your actions while applying AI."

Do you know AI can even help predict text and you can write complete stories using AI? Maybe this book itself is written using AI! Whether we recognize it or not, AI has already made its way into our lives. Predictive searches on Google, 'recommended' videos on platforms like YouTube or a software designed to prompt and correct text while typing, are all examples of Artificial Intelligence. Generating human-like outputs is one aspect of AI we have gained enough control over. Now, the natural question is, can AI also make human-like decisions?

Do you know that a machine learning algorithm **OpenAI's GPT-2** language model is trained to predict text? It takes months of training over tons of data on expensive computers but once that is done, it is easy to use.

Recently, AI systems have been developed to analyze images to distinguish between a benign skin lesion and melanoma cancer, matching the accuracy of 21 certified dermatologists. This will not only make up for the lack of dermatologists but also make their task simple and more efficient.

This example shows that AI can make good decisions, even better than humans. One thing to notice, however, is that an algorithm is more likely to label an image as cancer if there is a ruler in the image. Dermatologists use a ruler in the photo to measure the size of a skin lesion. Another issue is that with algorithms like this in place, people might not even consider seeing a dermatologist and may click a picture of the skin lesion using mobile phones and find out if it is anything to be worried about. However, the problem is that these photos are likely to differ from those of a dermatologist in terms of lighting, photo quality, zoom, etc. Simple factors like skin color or how hairy one's arm is can put the algorithm outside its original training conditions. And it is well known that when algorithms are given data that is different from what they were trained with, they can behave unpredictably.

Consider the case of self-driving cars. Imagine an automated car is fed with an algorithm to save the passengers sitting in the car over the people outside the car, in case of accidents, otherwise no one will buy the car. During a collision, to save one passenger sitting inside the car, the car runs over a group of six people. This makes a self-driving car socially unacceptable, especially to those who cannot afford to buy it.

You can use this tool to predict text from a data set of text:

Scan QR code or visit:
http://openai.com/blog/better-language-models/

So, if AI and machines are to take over in future, how can we make their decision making reliable? How can we make their decisions predictable and convincing under all circumstances? What will make these autonomous, self-improving, independent machines and software trustworthy?

This is where ethics come into the picture. Ethics are a loosely-defined set of moral principles about right and wrong guiding actions of an individual or a group. Since technology itself does not possess moral or ethical qualities, it needs to be fed with human ethics. When designed and tested well, it arrives at predictable outputs for predictable inputs via such a set of rules or decision paths.

But here are two challenges. First, how does the team of developers determine what is a good or right outcome and for whom? Is this outcome universally good or is it good

only for some? Is this outcome good under certain contexts or situations and not under other conditions? Is it good against certain standards but not good against others? These discussions, questions and answers 'chosen' by the team are critical. The second challenge is that AI is an autonomous, self-learning and self-improving technology. This means that it cannot be fed and does most of its decision-making itself based on its own analysis of data.

SUMMARY OF WHAT ALL IS COVERED UNDER ETHICS

Basics of AI Ethics:

1. Bias, Prejudice and Fairness
2. Accountability
3. Transparency, Interpretability and Explainability

Actions of AI:

1. Safety
2. Human-AI Interaction
3. Cybersecurity and Wrong Intentional Use
4. Data Privacy and Control

Impact of AI:

1. Job Losses and Unemployment
2. Civil Rights—Robot Rights
3. Human-Human Interaction Change
4. Economy

A fundamental debate is underway that AI will change the way our society works and it is very important to plan for such a society in advance before AI gets too involved in our lives.

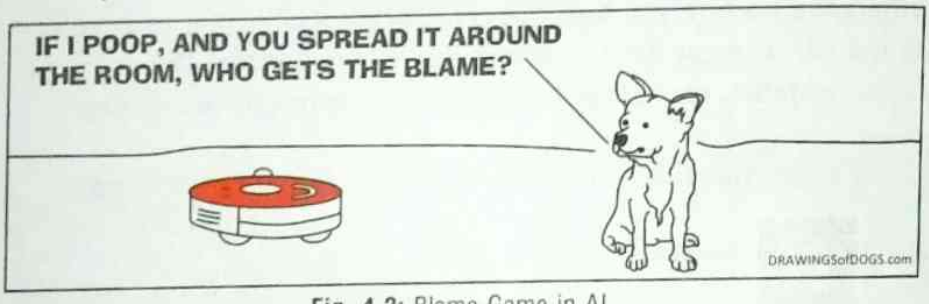


Fig. 4.2: Blame Game in AI

At the core of these two challenges lies the problem of existing human biases which will enter AI systems through both the developers and data. In the very process of its creation, technology becomes inherently biased against the people who create it. It exhibits the opinions, understanding and ethical stand of its creators. Thus, ethics of a technology start with the ethics of its creators.

"Unfortunately, we have biases that live in our data, and if we don't acknowledge that and if we don't take specific actions to address it then we're just going to continue to perpetuate them or even make them worse."

—Kathy Baxter, Ethical AI Practice Architect, Salesforce

Potential Harms from Automated Decision-making

Individual Harms		Collective/ Societal Harms
Illegal	**Unfair**	
Loss of Opportunity		
Employment Discrimination		
E.g., Filtering job candidates by race or genetic/health information	*E.g.*, Filtering candidates by work proximity leads to excluding minorities	**Differential Access to Job Opportunities**
Insurance & Social Benefit Discrimination		
Eg., Higher termination rate for benefit eligibility by religious group	*E.g.*, Increasing auto insurance prices for night-shift workers	**Differential Access to Insurance & Benefits**
Housing Discrimination		
Eg., Landlord relies on search results suggesting criminal history by race	*E.g.*, Matching algorithm less likely to provide suitable housing for minorities	**Differential Access to Housing**
Education Discrimination		
Eg., Denial of opportunity for a student in a certain ability category	*E.g.*, Presenting only ads on for-profit colleges to low-income individuals	**Differential Access to Education**
Economic Loss		
Credit Discrimination		
Eg., Denying credit to all residents in specified neighbourhoods (redlining)	*E.g.*, Not presenting certain credit offers to members of certain groups	**Differential Access to Credit**
Differential Pricing of Goods and Services		
Eg., Raising online prices based on membership in a protected class	*E.g.*, Presenting product discounts based on "ethnic affinity"	**Differential Access to Goods and Services**
	Narrowing of Choice	**Narrowing of Choice for Groups**
	E.g., Presenting ads based solely on past "clicks"	
Social Detriment		
	Network Bubbles	**Filter Bubbles**
	E.g., Varied exposure to opportunity or evaluation base on "who you know"	*E.g.*, Algorithms that promote only familiar news and information
	Dignitary Harms	**Stereotype Reinforcement**
	E.g., Emotional distress due to bias or a decision based on incorrect data	*E.g.*, Assumption that computed decisions are inherently unbiased
	Constraints of Bias	**Confirmation Bias**
	E.g., Constrained conceptions of career prospects based on search results	*E.g.*, All-male image search results for "CEO," all-female results for "teacher"
Loss of Liberty		
	Constraints of Suspicion	**Increased Surveillance**
	E.g., Emotional, dignitary and social impacts of increased surveillance	*E.g.*, Use of 'predictive policing' to police minority neighbourhoods more
Individual Incarceration		**Disproportionate Incarceration**
E.g., Use of "recidivism scores" to determine prison sentence length (legal status uncertain)		*E.g.*, Incarceration of groups at higher rates based on historic policing data

(Source: *Future of Privacy Forum*)

Fig. 4.3: Ethics-related Issues with AI

There are 104 cognitive biases that affect human decision-making. And, hence, the AI we build. AI can amplify human biases by virtue of how it learns and the feedback it gets. Such biased algorithmic systems can lead to undesirable and unfair outcomes, unequal and unjust consequences. And the intensity of these implications can be very severe depending on the type and number of people it affects.

Understanding bias in AI starts with tracing its sources and then identifying which of those can be countered with technology. Bias can be of two types—Model bias, which means the model is able to adequately represent the data set accurately enough, and the other bias being Prejudice bias, where the creator's stand on certain issues is embedded intentionally or unintentionally in the algorithm.

Let us understand this with a simple activity. Grab a sheet of paper and plot some points on it. Now, try to draw a line through this random distribution of points. More widely spread apart the points, curvier the line. If you continue to add more points, the line no longer fits all the points. Now you start looking for the "best possible" curve fit. In Mathematics, your line can be represented by one equation (or function) and "the best possible" calculation by another function. This is the fundamental idea behind machine learning model.

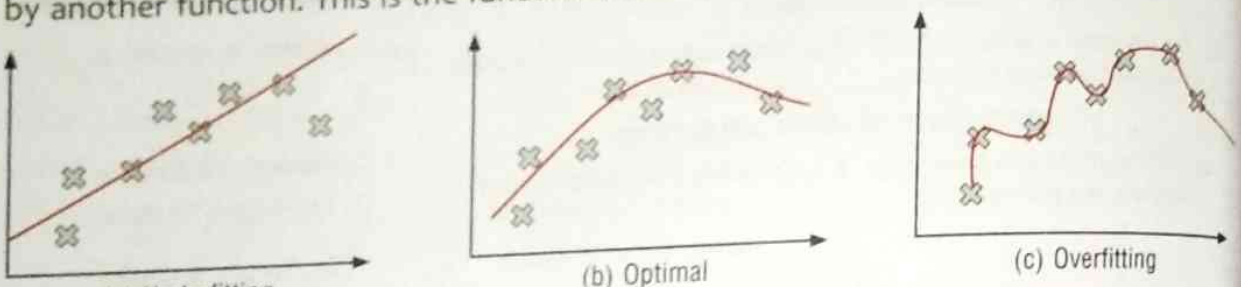


Fig. 4.4: Curve Fitting

List of cognitive biases:

Scan QR code or visit:
https://traviswhitecommunications.com/wp-content/uploads/2018/01/Cognitive-Biases-Grouped.png

As you can see in Fig. 4.4, if you try to put the curve through maximum number of points, you end up having a vague representation and it will be very difficult to extrapolate and predict the nature of curve for unseen data. In fact, the equation will be very difficult to establish for such a random data. Any close forced fit will lead to underrepresentation of minority data which leads to biases. Whatever function y=f(x) we try to establish with the data, there will always be a set of new data for which the function does not hold true, so you end up making some assumptions which are inherent biases.

Consider an example of a facial emotion recognition AI model trained on faces from around the world. Here, we have assumed that we have enough diversity in terms of age, gender, race, ethnicity, etc. We all know that the intensity with which we express

our emotions is also correlated with other qualitative factors. For example, the size of smile is correlated with what is cultural appropriateness, how deeply people relate to the trigger, how they feel about their teeth, how often they were criticized for laughing too loudly, their mental state at that moment, etc. What happens if we try to use this model to assess how engaged and happy students are in a classroom? Can we have a function to represent all possible versions of happiness on the basis of these considerations using visible and measurable features like size of smile, volume of laughter, how wide the eyes open? What do you think?

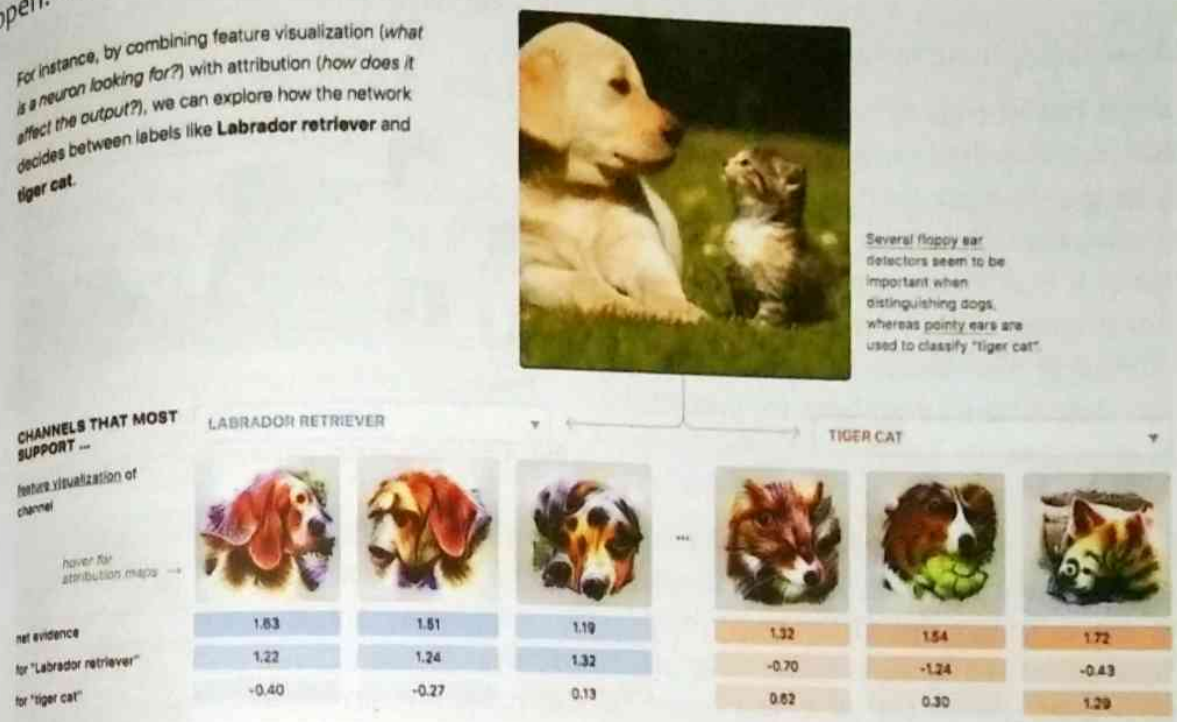


Fig. 4.5: What is a neural network looking for and how is it attributing what it sees? (Source: *https://distill.pub/2018/building-blocks/*)

As you can observe, historical bias already exists in data. Depending on how the data set is created, Representation Bias and Measurement Bias also creep in. Evaluation and Aggregation biases come into play depending upon the model we deploy.

Let us now discuss accountability, safety and fairness in AI.

"There's a real danger of systematizing the discrimination we have in society [through AI technologies]. What I think we need to do—as we're moving into this world full of invisible algorithms everywhere—is that we have to be very explicit, or have a disclaimer, about what our error rates are like."

—Timnit Gebru, Research Scientist, Google AI

Accountability in AI may be achieved by human audits, impact assessments or via governance through policy or regulation. Tech companies generally prefer self-regulation but now they realize the need for external intervention. Governance through 'human-in-the-loop', where certain decisions identified as high-risk require vetting by a human, have also been proposed as a model for accountability.

Safety in AI can be understood as not causing accidents or exhibiting unintended or harmful behaviour. Harm to humans is obvious and safety concerns with autonomous vehicles, drone deliveries are well known. Did you read about the recent attack in Saudi Arabia by an Iranian drone? Such is the power of a small drone. How can we model and enable safety in an autonomous system?

In a rule-based engine, in which certain inputs result in a specified output, safety measures can be put in place with testing. With AI, as the complications go far beyond simple rule engine, it is difficult to assess all scenarios. Autonomous decision-making requires automating the ability to evaluate safety under uncertain conditions to predictably prevent harm.

Fig. 4.6: Containment Strategy (Source: *Iyad Rahwan, MIT*)

"Fairness is a big issue. Human behaviour is already discriminatory in many respects. The data we've accumulated is discriminatory. How can we use technology and AI to reduce discrimination and increase fairness? There are interesting works around adversarial neural networks and different technologies that we can use to bias toward fairness, rather than perpetuate the discrimination. I think we're in an era where responsibility is something you need to design and think about as we're putting these new systems out there so we don't have these adverse outcomes."

—Paul Daugherty, Chief Technology and Innovation Officer, Accenture

ANTI-BULLYING WITH AI

Bullying is something that many of us might have encountered at some point in our life or at least read about. You should never bully anyone, help others who might be bullied and support an anti-bullying environment in your school at all times.

It is very important to recognize when bullying is happening. Sometimes, we may say something which we consider to be harmless but it may offend the other person and they may feel bullied depending on their experiences, culture, religion or family values. You should be very careful not to hurt anyone's sentiments and beliefs.

Sometimes, you might post something on social media or send an email which might contain some words or phrases which could be interpreted as offensive or bullying in nature. At times, even typos, grammatical errors or sarcasm may be misunderstood as bullying.

Can AI help?

Yes, absolutely.

Let us look at the following AI tool:

https://mycomputerbrain.net/php/experiments/ai.experiment19b.php

It helps you train an AI application to recognize texts/posts and match them to a particular output. No coding as such is required. You can enter a total of 12 expressions, including single words and sentences.

The instructions are on the right side of the menu and help you to:

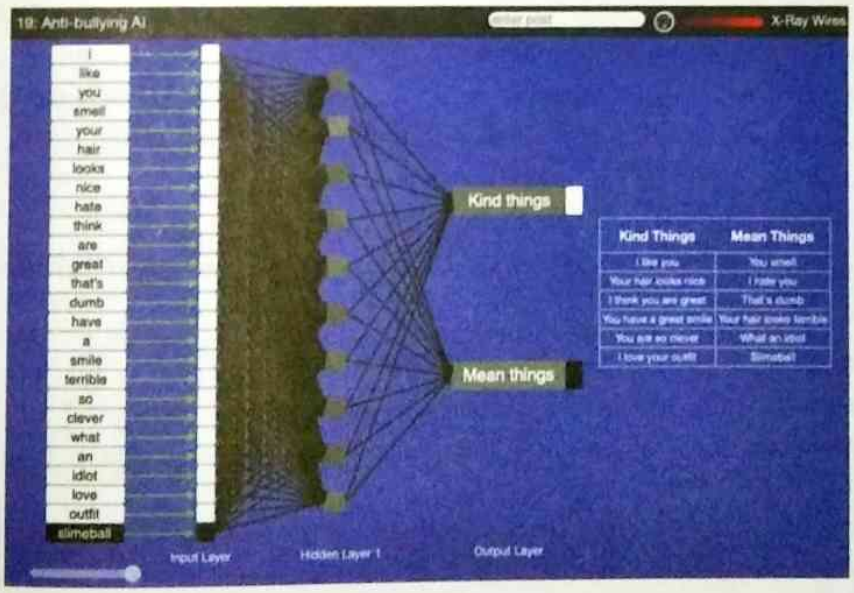
- Create the training data
- Train the Artificial Neural Network (ANN) through the **Start Learning button** (it will say **Learning Completed** when finished)
- Test the ANN by entering short sentences and phrases

On entering the inputs, ANN will be constructed in real time so you can analyze how AI is giving the output in real life.

Note: *Your input is not stored. When the browser window is refreshed, the experiment is reset. So you can play with it as many times as you want.*

Discuss your learning with your friends and come up with your own training data to test it further.

- How does AI decide when good words and when mean words are used in combination?
- Can AI pick up sarcasm in phrases?
- Can you improve the AI model further?



This is an example table with expressions that most people would consider either "kind" (left-hand column) or "mean" (right-hand column).

The grey boxes in the input, hidden and output layer are called 'perceptrons'. We will learn more about perceptron model in Chapter 6 when we study Neural Networks.

A perceptron is a model or program that simulates human brain intelligence. On the left-hand side of the ANN, we see all words and expressions in the table. They are provided to the ANN as training data. The ANN calculates whether to classify them into the "Kind things" or the "Mean things" category based upon user data.

Discuss with your friends how you can use such a tool while you are writing mails or using social media. Did it help you understand what is considered as bullying behaviour?

EXERCISES

Objective Type Questions

A. Fill in the blanks:

1. Human biases are magnified in AI due to the way it ..... and takes ..... .
2. Algorithms behave ..... when they are given data other than what they are trained with.
3. Technology loses its neutral stand at ..... .
4. Minority or unique features get ..... attention in machine learning models.
5. In a rules-based system, safety can be ensured by testing and ..... procedures.
6. Bias can enter AI through ..... and ..... .
7. Ethics of technology start with ..... .
8. Data accumulated is ..... because human behaviour is already discriminatory.

B. State whether the following statements are True or False.

1. AI can be trained to write a story from few lines of text.
2. Presenting only ads on for-profit colleges to low-income individuals is an example of education discrimination.
3. Self-driving cars are free from any liabilities and regulations as no human is driving the car.
4. AI biases are as good as human biases involved in algorithm.

5. All-male image search results for "CEO," all-female results for "teacher" is an example of confirmation bias.
6. Bias in AI can be reduced with more training data.
7. Use of "predictive policing" to police minority neighbourhoods more, is inherently a bias causing loss of liberty.
8. AI algorithms reading your emails to suggest you responses is not an example of data protection and privacy concern.

C. Multiple Choice Questions (MCQs):

1. Machine learning understands data through:
   (a) developer (b) set of codes
   (c) ethics (d) pattern
2. How many types of cognitive biases are there that impact human decision-making?
   (a) 108 (b) 106
   (c) 104 (d) 102
3. Which of the following are instances of gender bias?
   (a) digital assistance having female voice
   (b) searching 'hands' on internet are white hands
   (c) women candidate selected as 'chef' by AI system
   (d) 'o bir doktor' from Turkish translated as 'he is a doctor'
4. Which of the following is untrue about AI?
   (a) AI is inherently biased.
   (b) AI can produce predictable output for well-designed set of rules and decision paths.
   (c) AI takes up the bias of the developer.
   (d) AI is unpredictable under strange situations.
5. Which one of the following is not an area of AI?
   (a) computer vision (b) voice recognition
   (c) web design (d) image recognition
6. Which of these does NOT use machine learning/AI?
   (a) driverless cars (b) SIRI/Alexa
   (c) Sonos wireless speakers (d) facial recognition on your phone
7. Which of the following is not a goal of AI?
   (a) fairness (b) accountability
   (c) transparency (d) obligation
8. Which algorithm is used to predict text and can be used in book writing as well?
   (a) Forest Star (b) Alpha Pro
   (c) Random function (d) OpenAI's GPT-2

Subjective Type Questions

1. If a machine based on AI makes a decision with unintended consequences, who is responsible?
2. An online search shows mostly male candidates when searched for 'doctors' and female candidates when searched for 'nurses'. Should the developer be held accountable for this bias?

3. A company decides to select candidates using an AI system for the role of 'Head Chef'. What all biases are likely to be present in this case?
4. In a company, visitors enter their personal details in visitors' book. However, the company shares this data with some marketing companies. Do you think it is ethical? Why/Why not?
5. Data privacy is a serious concern when it comes to Artificial Intelligence. What steps are taken by various countries to deal with this problem?
6. A self-driving car is programmed to protect innocents from criminals. In case of accidents, it hits a murderer on the road rather than a puppy. Should the developer be held accountable for this decision?
7. Some experts say robots for military combats will be more efficient. However, others challenge it on ethical grounds. What are the ethical challenges to it?
8. Human values are simple at its core. However, factors like culture, law, religion, etc., make them complex. This complexity is at the core of ethics of AI. Discuss with examples.
9. Do you consider it ethical to incorporate the use of AI for various jobs, given that it takes away human jobs and magnifies human bias?
10. How do you think income would be shared if AI is used in place of human workforce?
11. AI will probably bring with it many health benefits and learning opportunities. How will these benefits be made accessible and available to all the people in the society?
12. AI is a powerful tool in various fields; however, depending on how it is used, it can either be a blessing or a curse. Discuss with an example.
13. How can learning opportunities for AI be extended to all?
14. How will human beings ensure that they stay ahead of Artificial Intelligence?

Quick Activity

Imagine a major landslide occurs in a landslide-prone area in which around 20,000 people are trapped, including pilgrims, tourists and other locals. The trapped population comprises women, children, senior citizens, patients in hospitals, political leaders, prisoners in jail, etc. If you get to design an AI system to aid and rescue in such scenarios, what priority order will you set for the rescue operation?

Fun Time

- See how Oxford University is working on an institute of AI.

Scan QR code or visit:
https://www.youtube.com/watch?v=5JVFKrgmr-8

What do you think about the role of such an institute? Can you think of something like this in India? Prepare a presentation for such an institute and explain how it will work.