6CCS3AIN

Week 10: AI and Ethics: Tutorial Questions

Q1. In the lectures, we mention that for model-driven (or symbolic) AI systems it is usually straightforward to generate explanations of the reasoning used by the AI system in each particular case where the system is applied. An expert system which uses IF-THEN rules is a common example of a model-driven AI system. Why is it usually straightforward to generate an explanation? How are the explanations generated?

Q2. The technology company you work for has tasked you with developing an AI application which searches social media, such as Facebook and Linked-In, for information and photos of potential recruits and then matches this information against profiles of the company's best-performing existing staff to identify the best potential recruits. The plan is that the Human Resources Department will contact the potential recruits to invite them for an interview.

After developing the system but before putting it into production, you notice that almost all the recommended potential recruits are men. You realize that this may be because most of the staff in the technology sector, including most of company's existing staff and most potential recruits, are men.

You also notice that the system seems to reject any potential recruits whose photos show them wearing a hat or other headgear. You do not know why the AI system does this, but it may just be some trivial quirk of a machine learning system.

You have several possible action options:

(a) Do nothing. All AI systems have quirks, and it is best to leave them alone.

(b) This decision is the responsibility of my employer, so I will just follow the orders of my boss.

(c) I will try to eliminate the bias against both women and hat-wearers.

(d) I will try to eliminate the bias against women but ignore the issue of the hat-wearers, as this is trivial.

(e) I will try to eliminate the bias against hat-wearers but ignore the issue of gender bias, because this is a problem across the entire technology sector which one company cannot solve.

What are the reasons for or against each option? Which options would you definitely not choose? Which option would you choose?

Q3. In the KEATS Quiz for Week 9, there is a question about a flag-colouring problem with 5 agents (the question is called “Pentagon”). This is the question:

> We have a distributed system with 5 agents. Each agent is connected to two others, arranged in a circle (ie, forming the corners of a pentagon).
>
> Each agent has two possible colours (or states), Red or Blue. The initial state is that one agent is Blue and the other 4 agents are Red.
>
> Each agent has been programmed with the following algorithm:
>
> START ALGO
>
> To decide the colour it will take at the next round, each agent looks to its two neighbours and executes the following two rules in order.
>
> Rule 1: If the two neighbours are both currently the same colour as each other, then the agent chooses the opposite colour to its two neighbours. It does this without regard to its own current colour.
>
> Rule 2: If the two neighbours are different colours from each other, then the agent remains whatever colour it currently is.
>
> END ALGO
>
> Which of the following statements are true?

There is one correct answer in the quiz to this question, which is a statement about the behaviour of the model. Provide an explanation for the reasoning used to reach this correct statement.

6CCS3AIN

Week 10: AI and Ethics: Tutorial Questions and Solutions

Q1. In the lectures, we mention that for model-driven (or symbolic) AI systems it is usually straightforward to generate explanations of the reasoning used by the AI system in each particular case where the system is applied. An expert system which uses IF-THEN rules is a common example of a model-driven AI system. Why is it usually straightforward to generate an explanation? How are the explanations generated?

Q1/SOLUTION

It is usually straightforward because we can trace through the reasoning process used by the AI system itself. In the case of an expert system with IF-THEN rules, we can see which specific rules were invoked (used) by the system to reach the particular decision or recommendation it reached in any particular case. Such a sequence of specific rules is called a *trace* of the system.

Likewise, when we reason over a graph (as in the Voter Model graph in Q3 in these Exercises), we can trace through the particular sequence of deductions we have drawn to reach a case. When we have a model of a domain, the model normally allows us to do this.

Q2. The technology company you work for has tasked you with developing an AI application which searches social media, such as Facebook and Linked-In, for information and photos of potential recruits and then matches this information against profiles of the company's best-performing existing staff to identify the best potential recruits. The plan is that the Human Resources Department will contact the potential recruits to invite them for an interview.

After developing the system but before putting it into production, you notice that almost all the recommended potential recruits are men. You realize that this may be because most of the staff in the technology sector, including most of company's existing staff and most potential recruits, are men.

You also notice that the system seems to reject any potential recruits whose photos show them wearing a hat or other headgear. You do not know why the AI system does this, but it may just be some trivial quirk of a machine learning system.

You have several possible action options:

(a) Do nothing. All AI systems have quirks, and it is best to leave them alone.

(b) This decision is the responsibility of my employer, so I will just follow the orders of my boss.

(c) I will try to eliminate the bias against both women and hat-wearers.

(d) I will try to eliminate the bias against women but ignore the issue of the hat-wearers, as this is trivial.

(e) I will try to eliminate the bias against hat-wearers but ignore the issue of gender bias, because this is a problem across the entire technology sector which one company cannot solve.

What are the reasons for or against each option? Which options would you definitely not choose? Which option would you choose?

Q2/SOLUTION

For this question, first think through all the positives and negatives of each option in turn.

For example, for option (a), a positive of Doing Nothing is that it is easy to do. A negative is that there is a risk of the company being accused of discrimination for which there may severe legal penalties, particularly if the company or its employees knew about the problem but did nothing.

Options you should NOT choose are to do nothing, or to blindly follow orders. Doing either of these may well lead legal or other consequences to you.

Hat-wearing may be a proxy variable for something else, such as religious affiliation. So, we should not dismiss it as trivial until it has been investigated. Option (d) is therefore not advisable to take.

Discrimination on the basis of gender is illegal in most developed countries, so option (e) is also not advisable to take.

Of these options, the only one which is advisable to select is option (c), to try to eliminate the bias against both gender and hat-wearing.

Q3. In the KEATS Quiz for Week 9, there is a question about a flag-colouring problem with 5 agents (the question is called “Pentagon”). This is the question:

> We have a distributed system with 5 agents. Each agent is connected to two others, arranged in a circle (ie, forming the corners of a pentagon).
>
> Each agent has two possible colours (or states), Red or Blue. The initial state is that one agent is Blue and the other 4 agents are Red.
>
> Each agent has been programmed with the following algorithm:
>
> START ALGO
>
> To decide the colour it will take at the next round, each agent looks to its two neighbours and executes the following two rules in order.
>
> Rule 1: If the two neighbours are both currently the same colour as each other, then the agent chooses the opposite colour to its two neighbours. It does this without regard to its own current colour.
>
> Rule 2: If the two neighbours are different colours from each other, then the agent remains whatever colour it currently is.
>
> END ALGO
>
> Which of the following statements are true?

There is one correct answer in the quiz to this question, which is a statement about the behaviour of the model. Provide an explanation for the reasoning used to reach this correct statement.

Q3/SOLUTION

1. Start by drawing the graph of the problem in the initial state (initial configuration of colours).

2. Then, take each node in turn and apply the algorithm to deduce what colour that node will be in the second round. Take a note of your reasoning for each node, eg,

   "Let us take the Blue node first. Both its neighbours are Red. Then Rule 1 applies, and this node must become the opposite colour to Red, ie, it must become Blue. It is already Blue, so it stays Blue. Let us now take the Red node which is clockwise neighbor from the Blue node. This node has a Blue neighbor and a Red neighbor, so Rule 2 applies. So this node stays the same colour, ie Red. . . . ."

3. Repeat for several rounds. You should see that the system state reaches an equilibrium state of 2 blue and 3 red nodes. This steady state does not change.

4. Now write out the reasoning you went through in the above steps to generate an explanation.