6 ccs3ain

第 10 周:人工智能与伦理:辅导题

Q1。在课程中，我们提到，对于模型驱动(或符号)的人工智能系统，在每个特定的系统应用情况下，通常直接生成人工智能系统使用的推理解释。使用 IF-THEN 规则的专家系统是模型驱动 AI 系统的常见例子。为什么解释通常很简单?这些解释是如何产生的?

Q2。你工作的技术公司负责你与开发人工智能应用程序搜索社交媒体,比如 Facebook 和 linkedin,信息和潜在的新兵的照片然后匹配这些信息对公司资料的最好表现最好的现有员工识别潜在的新兵。计划是人力资源部会联系潜在的应聘者，邀请他们参加面试。

在开发系统之后，但在投入生产之前，您会注意到几乎所有推荐的潜在员工都是男性。你会意识到，这可能是因为科技行业的大多数员工，包括公司现有的员工和潜在的新员工，都是男性。

你还会注意到，系统似乎会拒绝任何在照片上显示戴着帽子或其他帽子的潜在新兵。你不知道为什么人工智能系统会这样做，但这可能只是机器学习系统的一个小怪癖。

你有几个可能的行动选项:

(一)什么都不做。所有 AI 系统都有怪癖，最好不要去管它们。

这个决定是我老板的责任，所以我只会服从老板的命令。

(c)我将努力消除对妇女和戴帽子者的偏见。

(d)我将努力消除对妇女的偏见，但忽略戴帽子者的问题，因为这是微不足道的。

(e)我会努力消除对戴帽子者的偏见，但忽略性别偏见的问题，因为这是一个横跨整个技术部门的问题，一个公司无法解决。

支持或反对每一种选择的理由是什么?选择呢?你会选择哪一个?　　哪些选择是你绝对不会选择的

第三季。在第 9 周的济慈测验中，有一个问题是关于五个代理人的国旗颜色问题(这个问题叫做“五角大楼”)。这是一个问题:

我们有一个分布式系统，有 5 个代理。每个代理都与另外两个连接在一起，排成一个圆圈(即五边形的四角)。每个代理有两种可能的颜色(或状态)，红色或蓝色。初始状态是，一个代理是蓝色的，其他 4 个代理是红色的。每个 agent 的编程算法如下:

开始算法

为了决定下一轮它将采用的颜色，每个代理观察它的两个邻居，并执行以下两个规则顺序。

规则 1:如果这两个邻居目前都是相同的颜色，那么代理选择相反的颜色邻居。它这样做时并不考虑自己当前的颜色。

规则 2:如果两个邻居的颜色不同，那么代理保持当前的颜色。

算法结束

下列哪个陈述是正确的?

在这个问题的测试中有一个正确的答案，这是一个关于模型行为的陈述。请解释得出这一正确结论的原因。

6 ccs3ain

第 10 周:人工智能和伦理:教程问题和解决方案

Q1。在课程中，我们提到，对于模型驱动(或符号)的人工智能系统，在每个特定的系统应用情况下，通常直接生成人工智能系统使用的推理解释。使用 IF-THEN 规则的专家系统是模型驱动 AI 系统的常见例子。为什么解释通常很简单?这些解释是如何产生的?

Q1 /解决方案

它通常是直接的，因为我们可以追踪 AI 系统本身使用的推理过程。在具有 IF-THEN 规则的专家系统的情况下，我们可以看到系统调用(使用)了哪些特定的规则，以达成它在任何特定情况下达成的特定决策或建议。这种特定规则的序列称为系统的跟踪。

同样，当我们对一个图进行推理时(如本练习 Q3 中的选民模型图)，我们可以通过我们所绘制的特定推理序列来得出一个案例。当我们有一个域的模型时，这个模型通常允许我们这样做。

Q2。你工作的技术公司负责你与开发人工智能应用程序搜索社交媒体,比如 Facebook 和 linkedin,信息和潜在的新兵的照片然后匹配这些信息对公司资料的最好表现最好的现有员工识别潜在的新兵。计划是人力资源部会联系潜在的应聘者，邀请他们参加面试。

在开发系统之后，但在投入生产之前，您会注意到几乎所有推荐的潜在员工都是男性。你会意识到，这可能是因为科技行业的大多数员工，包括公司现有的员工和潜在的新员工，都是男性。

你还会注意到，系统似乎会拒绝任何在照片上显示戴着帽子或其他帽子的潜在新兵。你不知道为什么人工智能系统会这样做，但这可能只是机器学习系统的一个小怪癖。

你有几个可能的行动选项:

(一)什么都不做。所有 AI 系统都有怪癖，最好不要去管它们。

这个决定是我老板的责任，所以我只会服从老板的命令。

(c)我将努力消除对妇女和戴帽子者的偏见。

(d)我将努力消除对妇女的偏见，但忽略戴帽子者的问题，因为这是微不足道的。

(e)我会努力消除对戴帽子者的偏见，但忽略性别偏见的问题，因为这是一个横跨整个技术部门的问题，一个公司无法解决。

支持或反对每一种选择的理由是什么?选择呢?你会选择哪一个?　　哪些选择是你绝对不会选择的

Q2 /解决方案

对于这个问题，首先依次考虑每个选项的所有优点和缺点。

例如，对于选项(a)，“什么都不做”的一个积极因素是它很容易做。消极的一面是，公司可能会被指控歧视，可能会受到严厉的法律惩罚，特别是如果公司或员工知道问题，但什么都不做。你不应该选择什么都不做，或者盲目服从命令。做这两件事中的任何一件都可能给你带来法律或其他后果。

戴帽子可能是其他事情的代理变量，比如宗教信仰。因此，在调查之前，我们不应该认为这是微不足道的。因此，不建议采取(d)办法。

基于性别的歧视在大多数发达国家是非法的，所以选择(e)也是不可取的。

在这些选项中，唯一可取的选项是选项(c)，以设法消除对性别和戴帽子的偏见。

第三季。在第 9 周的济慈测验中，有一个问题是关于五个代理人的国旗颜色问题(这个问题叫做“五角大楼”)。这是一个问题:

> 我们有一个分布式系统，有 5 个代理。每个代理都被连接
> 另外两个，排成一个圆圈(即五边形的四角)。
>
> 每个代理有两种可能的颜色(或状态)，红色或蓝色。 最初的
> 状态是，一个是蓝的，另外 4 个是红的。
>
> 每个 agent 的编程算法如下:
>
> 开始算法
>
> 为了决定下一轮将采用的颜色，每个代理商要看自己的
> 两个邻接并按以下两条规则执行。
>
> 规则 1:如果两个邻居现在都是相同的颜色
> 另一种颜色，代理选择与它的两种颜色相反的颜色
> 邻居。它这样做时并不考虑自己当前的颜色。
>
> 规则 2:如果两个邻居的颜色不同，那么
> 代理保持当前的颜色。
>
> 算法结束
>
> 下列哪个陈述是正确的?

在这个问题的测试中有一个正确的答案，这是一个关于模型行为的陈述。请解释得出这一正确结论的原因。

第三季度/解决
方案

1.首先画出问题初始状态的图(颜色的初始配置)。

2.然后，依次取每个节点，并应用该算法来推断该节点在第二轮中的颜色。记下你对每个节点的推理，例如，

“让我们先看蓝色节点。它的两个邻国都是红色。然后规则 1 适用，这个节点必须变成与红色相反的颜色，也就是说，它必须变成蓝色。它已经是蓝色的，所以它保持蓝色。现在我们取红节点它是蓝节点顺时针方向的邻居。这个节点有一个蓝色邻居和一个红色邻居，因此适用规则 2。所以这个节点保持相同的颜色，即红色. . . . ."

3.重复几轮。你应该看到系统达到了 2 个蓝色节点和 3 个红色节点的平衡状态。这个稳态不变。

4.现在写出你在上面步骤中所做的推理来生成一个解释。