Bài Tâp AI tuần 04

1. Nêu ít nhất 2 ví dụ về việc tốt và xấu theo theo **consequentialism**? Lập luận tại sao bạn cho rằng như vậy.

Ví dụ 1: Robin Hood là một nhân vật anh hùng trong văn học dân gian của Anh, một cung thủ và kiếm thủ cừ khôi, ông được biết đến với việc "**cướp của người giàu chia cho người nghèo**". Mặc dù biết việc trộm cướp là xấu, trái vs pháp luật và đạo đức, nhưng với việc làm của ông thì việc trộm của những kẻ giàu và quan lại, tham ô để chia cho người dân nghèo thì nó là 1 việc làm tốt đối với người dân.

Ví dụ 2: Trong việc giáo dục con trẻ, việc áp dụng các biện pháp dạy dỗ như: la mắng, đánh phạt,… Trong 1 số trường hợp thì nó có thể gây ảnh hưởng đến tâm lý của trẻ nhỏ, không ít cũng nhiều, nhưng ngược lại thì một số trẻ nhỏ thì nhận thức được việc làm sai của chúng và sửa lỗi. Vậy việc dạy dỗ trẻ em bằng các biện pháp như: la mắng, đánh phát,… đều có mặt tốt và mặt xấu riêng của nó.

1. Nêu quan điểm của bạn về (có thể chọn 1 trong 2 hoặc chọn cả 2):  
   a. Việc đi học trực tiếp trong mùa dịch  
   b. Việc giáo viên bắt buộc làm bài tập hàng tuần  
   Đó là việc tốt hay xấu theo **golden rule**? Lập luận tại sao bạn cho rằng như vậy (viết dưới 500 từ cho 1 chủ đề).
2. Theo em đây là biện pháp tốt và nên áp dụng trong mùa dịch này nhằm đảm bảo các biện pháp an toàn cho học sinh, sinh viên. Mặc dù việc học cũng gặp nhiều khó khăn như trong việc trao đổi kiến thức giữa các học sinh, sinh viên cũng như các vấn đề mà nhiều bạn ngại đặt vấn đề cho giáo viên trước mọi người. Hoặc do vấn đề đường truyền mạng hay điện,… Nhưng đây là biện pháp tốt nên làm trong thời dịch hiện nay.

*Ghi chú: Đây là bài luận nêu quan điểm cá nhân, GV chỉ chấm về mức độ hiểu golden rule và suy luận logic, không chấm đúng sai quan điểm của bạn.*

1. **Explainable AI** (XAI) là gì? Ví dụ các thư viện XAI.

XAI là một thuật toán trong AI, chuyên dùng để phân tích, giải thích các hoạt động của AI. Trả lời các câu hỏi vì sao AI lại đưa ra quyết định như vậy.

Explainable AI (XAI) là những thuật toán trong AI mà có thể truy trách nhiệm khi cần thiết.Ta có thể tạm dịch là AI có thể giải thích được là chìa khóa giải đáp thắc mắc. Nó lí giải các quyết định của hệ thống AI, cố gắng hiểu các bước liên quan để đưa ra quyết định.

Ví dụ các thư viện XAI:

SHAP: giải thích dự đoán của trường hợp bằng cách xác định mỗi tính năng đã đóng góp bao nhiêu vào dự đoán cuối cùng.   
LIME: giải thích dự đoán của mô hình về một cá thể bằng cách ước lượng mô hình với một mô hình có thể diễn giải về bản chất như cây quyết định.

NEO: Giống như SHAP và LIME, các neo tiếp cận một chiến lược tương tự để tạo ra các giải thích cục bộ. “ Nhưng thay vì các mô hình có thể diễn giải được sử dụng bởi LIME, các giải thích được thể hiện dưới dạng các quy tắc IF-THEN đơn giản, dễ hiểu ” .

1. Cho 2 **reward system** cho agent tìm đường thoát khỏi mê cung như bên dưới. Reward system nào tốt hơn? Tại sao?  
   System 1: +10 khi thoát khỏi mê cung, 0 khi di chuyển 1 bước trong mê cung  
   System 2: 0 khi thoát khỏi mê cung, -10 khi di chuyển 1 bước trong mê cung

System 1 sẽ tốt hơn. Vì khi hoạt động system 1, agent sẽ thoải mái hơn trong việc tìm lối ra mà không sợ bị mất điểm và agent sẽ cố tìm lối ra nhanh hơn, còn ở system 2, agent sẽ hoàn toàn bị hạn chế trong viêc cố gắng tìm đường ra, nhưng khi thoát được mê cung thì không được cộng điểm.

1. Thiết kế **reward system** cho agent tìm đường thoát khỏi mê cung mà bạn thấy tốt hơn 2 reward systems trong bài tập trên. Giải thích tại sao nó tốt hơn.

System : +10 khi thoát khỏi mê cung, 0 khi di chuyển 1 bước trong mê cung, -2 khi đi vào đường cục.

+ Khi thoát khỏi mê cung +10 cho agent giúp cho agent có thêm động lực và agent có thể cố gắng tìm đường ra

+ -2 khi đi vào đường cục sẽ làm hạn chế việc agent cố đi mà không hề tính toán điều gì, một số agent có thói quen bước lại vào đường cục cũ mà mình đã đi.