**Đạo đức AI**

Ai có thể gây ra những vấn đề đạo đức nào?

Sự thiên vị thuật toán và vi phạm quyền riêng tư

Trách nhiệm giải trình và quyền tự chủ

* Các khịa cạnh đạo đức của AI ngày càng sâu sắc và đa dạng

Agents : AI được xây dựng để có trí tuệ, hành động như con người, hành động hợp lý.

* có thể mở rộng những gì máy tính làm thành bộ dữ liệu lớn
* làm nhanh gần hư thời gian thực
* không mắc lỗi
* những hệ thống như này gọi là Agents

**What is agents ?**

Một tác nhân là bất kì cái gì có khả năng nhận thức môi trường của nó thông qua các cảm biến (sensors) và tác động lên môi trường đó thông qua các bộ truyền động ( actuators )

Bất kì cái gì : con người, người máy, software robots, các bộ ổn nhiệt,…

Agents luôn có tương tác với môi trường ( môi trường ở đây là môi trường làm việc của nó )

Thu nhận thông tin, nhận thức từ môi trường làm việc ( percepts) từ các sensors. Từ đó dùng các actuators để tạo ra các hành động ( action) để tác động vào môi trường làm việc.

* Tác tử con người :
  + Cảm biến (sensors ) : mắt, tai, các cơ quan khác
  + Bộ truyền động ( actuators ) : tay chân,….
* Tác tử robot:
  + Cảm biến (sensors ) : các cảm biến nhiệt, camera, cảm biến tia hồng ngoại,…
  + Bộ truyền động ( actuators ) : các động cơ
* Tác tử phần mềm:
  + Cảm biến (sensors ) : thiết bị đầu vào ( bàn phím, con chuột, camera micro, …)
  + Bộ truyền động ( actuators ) : thiết bị đầu ra ( màn hình, loa, máy in,…)

**Mô tả hành vi của các tác tử**

Hàm tác tử : ánh xạ từ chuỗi nhận thức đến 1 hành đồn [ f: P🡪A ]

Chương trình tác tử :

* Thực hiện việc mà hàm tác tử phải làm
* Tác tử = kiến trúc ( các phần cứng ) + chương trình

Ví dụ : tác tử máy hút bụi

Nhận thức : vị trí ( A or B ) và mức độ sạch sẽ ( sạch or bẩn )

Hành động của tác tử máy hút bụi

Có chuỗi các nhận thức, ứng với mỗi cái máy sẽ có hành động tương ứng

Từ đó ta sẽ tạo ra hàm tác tử để thực hiện hành động đó theo chuỗi nhận thức. ( ví dụ : nếu có chuỗi nhận thức này thì hành động gì ? tiếp theo sẽ có chuỗi nhận thức khác thì hành động thế nào ? )

Từ hàm tác tử ta sẽ viết chương trình hoàn chỉnh.

Cung cấp cho 1 tác tử :

* Cung cấp môi trường làm việc
* Cung cấp sensors
* Cung cấp actuators

Tác tử hợp lý :

* Cung cấp môi trường làm việc
* Cung cấp sensors
* Cung cấp actuators
* Cung cấp thêm đánh giá hiệu quả ( hàm mục tiêu ) ( ví dụ : máy hút bụi sẽ tìm nơi còn bẩn gần nhất )

**Bản chất của môi trường ( sau này có phỏng vấn sẽ cho tình huống cụ thể, từ đó hỏi về P E A S)**

Bao gồm : ( 4 tiêu chí P E A S cho 1 hệ thống ) :

* **P**erformance measure : tiêu chí đánh giá hiệu quả hoạt động
* **E**nvironment : Môi trường xung quanh
* Agrent’s **A**ctuators : các bộ phận hành động
* Agrent’s **S**ensors : các bộ phận cảm biến

Bước đầu tiên trong việc thiết kế 1 tác tử : Mô tả PEAS một cách đầy đủ nhất có thể

Ví dụ : Thiết kế 1 tác tử lái xe taxi tự động

Đánh giá hiệu quả hoạt động

Làm sao để chúng ta đánh giá trình điều khiển tự động ? : đánh giá theo các yếu tố :

* Đến đúng đích
* Giảm thiểu mức tiêu thụ nhiên liệu
* Giảm thiểu thời gian hoặc chi phí chuyến đi
* Giảm thiểu vi phạm luật giao thông
* Tối đa hoá sự an toàn và thoải mái…

Môi trường

* Một chiếc taxi phải xử lý nhiều loại đường
* Đèn giao thông, các phương tiện khác, người đi bộ, động vật, công trình,….
* Tương tác với khách hàng

Cảm biến :

Máy quay, GPS, đồng hồ tốc độ, đồng hồ đo khoảng cách,….

Bộ truyền động ( hành động ) :

Bánh lái, chân ga, còi xe, đèn tín hiệu,…

* Khi hỏi cần gì để thiết kế 1 tác tử Agrent thì cần trả lời theo P E A S

**Bài tập**

Thiết kế theo PEAS cho tác tử chẩn đoán bệnh tiểu đường.

Đánh giá hiệu quả hoạt động :

Chính xác ( sai số ít nhất ), an toàn, tốc độ nhanh

Chi phí bỏ ra thấp

Môi trường :

Bệnh nhân

Nhân viên bệnh viện

Phòng bệnh nhân

Cảm biến :

Đầu vào : bàn phím, scan tài liệu ( lấy dữ liệu xét nghiệm của bệnh nhân )

Micro : thu nhận dữ liệu bằng giọng nói

Camera : quan sát gương mặt, nét mặt, hành vi

Các bộ phận hành động :

Màn hình

Máy in báo cáo

Thiết kế theo PEAS cho tác tử nhặt đồ vật :

Đánh giá hiệu quả hoạt động :

Chính xác ( đồ vật không bị rơi ), tốc độ nhanh

Chi phí bỏ ra thấp

Môi trường :

Công viên

Đường xá

Bãi biển

Khu dân cư

Con người

Đường đi

Cây cối, đồ vật trên đường ( đèn giao thông, đèn công viên ,… )

Nhà cửa

Cảm biến :

Camera, cảm biến nhiệt, ánh sáng, hồng ngoại : quan sát để nhận biết đối tượng, môi trường

Micro : thu nhận thông tin bằng giọng nói

Các bộ phận hành động :

Động cơ nhặt đồ

Bài tập tại lớp