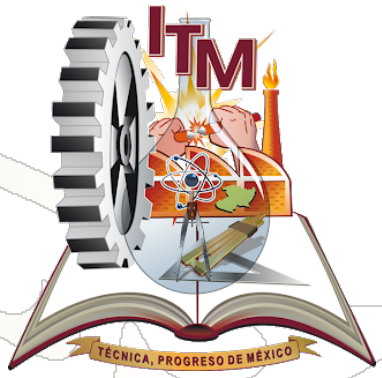


TECNOLOGICO NACIONAL DE MEXICO
INSTITUTO TECNOLOGICO DE MORELIA
"José María Morelos y Pavón"

Inteligencia Artificial

Ensayo inteligencia Artificial Un enfoque moderno (Capitulo 26 y 27)

Carlos Jahir Castro Cázares
17120151
Ingeniería en Sistemas Computacionales

03 de Marzo de 2021

1 Fundamentos Filosóficos

IA: ¿cómo trabaja la mente? ¿Es posible que las máquinas actúen de forma inteligente, igual que las personas? Y si así fuera, ¿tendrían mentes? ¿Cuáles son las implicaciones éticas de las máquinas inteligentes? Con estas preguntas empieza el capitulo 26 las cuales esperamos poder contestar o dar una respuesta a partir de los siguientes secciones. Primeramente, empezamos con las dos hipótesis de las IA.

IA débil
La IA consiste en la búsqueda del mejor programa agente en una arquitectura dada. Con esta premisa empieza la dada hipótesis de la IA débil, que difiere que una Inteligencia Artificial no puede ser inteligente si no que solo puede actuar inteligentemente, simulando la inteligencia verdadera para esto Alan Turing, en su famoso artículo Computing Machinery and Intelligence (Turing, 1950), sugirió que en vez de preguntar si las máquinas pueden pensar, deberíamos preguntar si las máquinas pueden aprobar un test de inteligencia conductiva (de comportamiento), conocido como el Test de Turing. Este test no se a podido pasar por alguna IA, pero se pensaba que esta puede ser superada para el año 2000, aumentando la capacidad de almacenamiento y procesamiento. Por otra parte, tenemos el **Argumento de la incapacidad**, aquí podemos ver el argumento que una maquina no puede realizar x cosas, como ser amable, tener recursos, ser guapo, etc. Pero si puede realizar tareas que para un humano serian muy pesadas y que requieren un grado de inteligencia, por lo cual no se puede juzgar una maquina por su capacidad de hacer o no hacer ciertas acciones. **La objeción matemática** es una afirmación que consiste en una maquina es inferior a un humano ya que esta tiene limitaciones por el teorema de la incompletitud y los hombres no estos atados a este teorema. Pero estas afirmaciones se pueden desmontar con varios argumentos que se explican mas claramente en el documento. **El argumento de la informalidad** una de las mas persistentes e influyentes criticas sobre la IA fue echa por Turing mediante el argumento de la informalidad del comportamiento, ya que en esencia la inteligencia huma es tan compleja como para poder ser representada por un conjunto de reglas que es en esencia lo que siguen las maquinas; Pero en contra parte tenemos la postura que se vino a llamar **Good Old-Fashioned AI** se supone que este término afirma que todo comportamiento inteligente puede ser capturado por un sistema que razona lógicamente a partir de un conjunto de hechos y reglas, los cuales describen el dominio. En resumen, los problemas que se le sacan a la inteligencia artificial a lo largo del tiempo son incorporaciones que se les dan a los diseños estándares de los agentes inteligentes y pude verse más como la evolución de la IA que más como limitaciones.

IA fuerte
Muchos filósofos han afirmado que una máquina que pasa el Test de Turing no quiere decir que esté realmente pensando, sería solamente una simulación de la acción de pensar. Con esto tenemos la otra hipótesis donde las maquinas si

piensan realmente. Para esto se han creado varios argumentos que son como el de la **consciencia** donde la maquina tiene que ser realmente consciente de sus propias acciones y estados mentales. Otros se centran en la **intencionalidad** esto es decir que esta tiene que tener la intención de realizar las acciones que realizara y no solamente seguir un conjunto de pasos. Con esto Turing sostuvo su argumento que las maquinas no pueden ser inteligentes. En 1848, Frederick Wöhler sintetizó urea artificial por primera vez. Este fue un hecho importante porque probó que la química orgánica y la inorgánica se podían unir, cuestión discutida muy fuertemente. Con esto podemos decir que la IA funciona en lo niveles mas bajos como la inteligencia orgánica pero no solo por ser iguales químicamente quiere decir que están al mismo nivel. Podemos concluir diciendo que en algunos casos el comportamiento de un artefacto es importante, aunque en otros sea el pedigrí del artefacto lo que importa. Lo importante en cada caso parece ser una cuestión de convención. **La teoría del funcionalismo** dice que un estado mental es cualquier condición causal inmediata entre la entrada y la salida. Bajo la teoría funcionalista, dos sistemas con procesos causales isomórficos tendrían los mismos estados mentales. Por tanto, un programa informático podría tener los mismos estados mentales que una persona. En contraste, la teoría del **naturalismo biológico** dice que los estados mentales son características emergentes de alto nivel originadas por procesos neurológicos de bajo nivel en las neuronas, y lo que importa son las propiedades de las neuronas. **El problema mente-cuerpo** cuestiona cómo se relacionan los estados y los procesos mentales con los estados y los procesos (específicamente del cerebro) del cuerpo. **El materialismo** mantiene que no existen los estados mentales tales como el sentir una sensación si no solo objetos materiales, pero esto debe de enfrentar por lo menos a los obstáculos de la **libertad de elección** en el como algo que solo se rige por las leyes de la física puede tomar elecciones como nosotros los humanos y el otro tiene que ver con el problema de la **conciencia** que sigue siendo un misterio hasta nuestros tiempos. Para comprobar estas hipótesis se han inventado varios experimentos hipotéticos como lo son el del **cerebro en una cubeta, la prótesis cerebral y la habitación china.** Que estos experimentos nos dan mas que pensar que respuestas claras por lo cual no son concluyentes.

La ética y los riesgos de desarrollar la Inteligencia Artificial
Hasta ahora nos hemos concentrado en si podemos desarrollar la IA, pero debemos también tener en cuenta si deberíamos hacerlo. La IA parece que expone problemas, por ejemplo:

- Las personas podrían perder sus trabajos por la automatización.
- Las personas podrían tener demasiado (o muy poco) tiempo de ocio.
- Las personas podrían perder el sentido de ser únicos.
- Las personas podrían perder algunos de sus derechos privados.

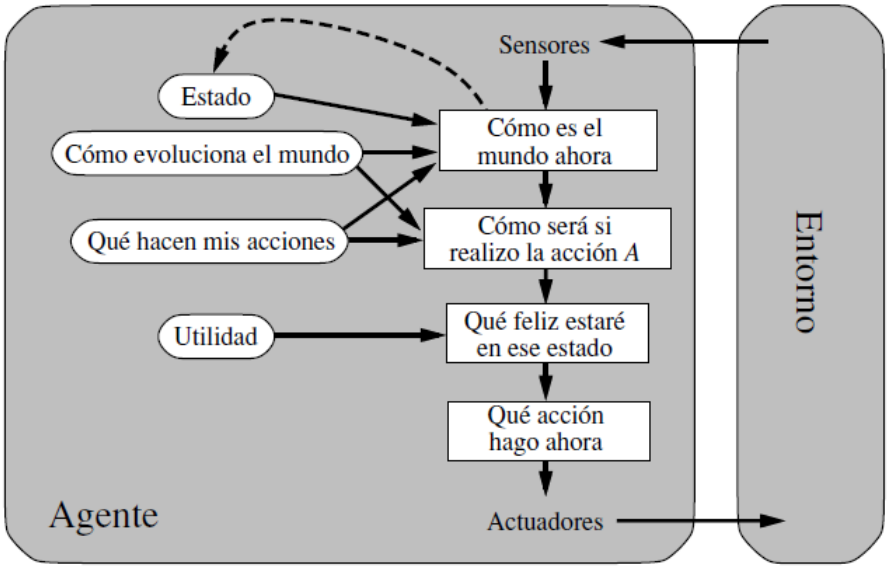
- La utilización de los sistemas de IA podría llevar a la pérdida de responsabilidad.
- El éxito de la IA podría significar el fin de la raza humana.
- Las personas podrían perder el sentido de ser únicos.

Otros riesgos que se pueden ver son la de la **singularidad tecnológica** donde la inteligencia de las maquinas superen a las humanas, y el **transhumanismo** donde llegue el momento en el que la distinción entre las maquinas y los humanos no sea lo suficientemente clara y termine con una fusión entre nosotros, por ultimo el dejar que los robots tengan conciencia lleva a debates como si es correcto usar como esclavos a criaturas consientes que nosotros creamos. Además, que si estos deberían de actuar moralmente o se les debería de programar el bien y el mal. Por ultimo se han identificado ocho amenazas potenciales para la sociedad que se exponen tanto ante la IA como ante una tecnología relacionada. Podemos concluir diciendo que algunas amenazas son improbables, pero merece la pena revisar dos de ellas en particular. La primera es que las máquinas ultra inteligentes podrían llevarnos a un futuro muy diferente del actual y puede que no sea de nuestro agrado. La segunda es que la tecnología de la robótica puede permitir a individuos con una psicopatía emplear armas de destrucción masiva. Concluimos diciendo que esto es más una amenaza de la biotecnología y nanotecnología que de la robótica.

2 IA: presente y futuro

Pueden existir diferentes diseños de agentes, desde agentes reactivos hasta agentes basados en conocimiento y completamente deliberativos. Además, los componentes de estos diseños pueden tener diferentes instanciaciones, por ejemplo, lógicas, probabilísticas o neuronales. Ha habido un tremendo avance tanto en el entendimiento científico como en nuestras habilidades tecnológicas en lo referente a los diseños y componentes de agentes.

Componentes de los agentes



Interacción con el entorno a través de sensores y actuadores: durante mucho tiempo en la historia de la IA, esto ha sido un notorio punto débil. Con unas pocas excepciones honorables, los sistemas IA se construyeron de tal forma que los humanos tenían que proporcionar las entradas e interpretar las salidas. Pero actualmente para los entornos físicos, entonces, los sistemas IA ya no tienen realmente excusa. Además, ya se dispone de un entorno enteramente nuevo como es Internet.

Seguir la pista del estado del mundo: ésta es una de las capacidades centrales que se requieren para un agente inteligente. Requiere tanto percepción como actualización de las representaciones internas. Las herramientas de filtrado se necesitan cuando está involucrada la percepción real (y por tanto imperfecta). Los algoritmos actuales de filtrado y de percepción pueden combinarse para hacer un trabajo razonable de informar predicados de bajo

nivel, pero se presenta un problema que es el de la incertidumbre de identidad que ha sido ignorado por las IA lógicas.

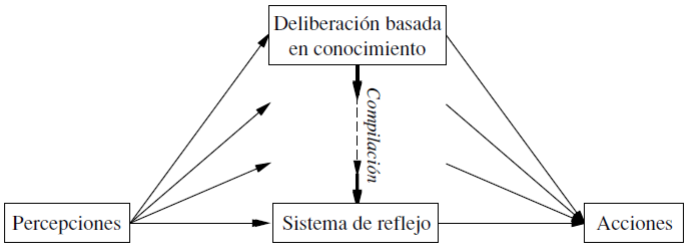
Proyección, evaluación y selección de cursos futuros de acción: los requisitos básicos de representación del conocimiento son los mismos aquí que para seguir la pista del mundo; la dificultad básica es hacer frente a los cursos de acción, tales como tener una conversación o tomar una taza de té, que finalmente constan de miles y millones de pasos primitivos para un agente real. Es sólo al imponer una estructura jerárquica que los humanos podemos abordarlos.

La utilidad como expresión de preferencias: en principio, basar las decisiones en la maximización de la utilidad esperada es completamente general y evita muchos de los problemas de los enfoques basados puramente en objetivos, tales como objetivos conflictivos y consecución incierta.

Aprendizaje: escriben la forma en que se puede formular el aprendizaje en un agente como aprendizaje inductivo (supervisado, sin supervisar o basado en el refuerzo) de las funciones que constituyen los diferentes componentes del agente.

Arquitecturas de agentes
Un agente completo debe ser capaz de hacer las dos cosas, utilizando una **arquitectura híbrida.** Una propiedad importante de las arquitecturas híbridas es que los límites entre los diferentes componentes de decisión no son fijos. **La IA en tiempo real** surgen medida que los sistemas IA entran en dominios más complejos, todos los problemas serán de tiempo real, porque el agente nunca tendrá tiempo suficiente como para resolver el problema de la decisión de forma exacta. En los últimos años han surgido dos técnicas prometedoras. La primera conlleva la utilización de **algoritmos de cualquier momento** un algoritmo de esta clase es un algoritmo cuya calidad de salida mejora gradualmente con el tiempo, de manera que tiene preparada una decisión razonable siempre que tenga una interrupción. La segunda técnica es el **metarazonamiento teórico para las decisiones** este método aplica la teoría del valor de la información, para la selección de cómputos. El valor de un cómputo depende tanto de sus costes como el de sus beneficios. El meta-razonamiento no es sino un aspecto de una **arquitectura reflexiva** general, es decir, una arquitectura que permite la deliberación sobre las entidades y las acciones computacionales que ocurren dentro de la misma arquitectura.

Componentes de los agentes



La utilidad como expresión de preferencias: Un agente perfectamente racional actúa en cualquier instante de tal manera que maximiza la utilidad esperada, dada la información que haya adquirido del entorno.

Racionalidad perfecta: Esta es la noción de la racionalidad que hemos utilizado implícitamente al diseñar agentes lógicos y teóricos para las decisiones. Un agente calculadoramente racional finalmente devuelve lo que habría sido la opción racional al comienzo de su deliberación.

Optimalidad limitada: Un agente óptimo limitado se comporta todo lo bien que puede, dado sus recursos computacionales. Es decir, la utilidad esperada del programa agente para un agente óptimo limitado es por lo menos tan elevada como la utilidad esperada de cualquier otro programa agente que se ejecute en la misma máquina.

3 Fundamentos matemáticos

Análisis de la complejidad y la notación O( )
Los científicos informáticos se suelen enfrentar con la tarea de comparar algoritmos para ver la rapidez de ejecución y la cantidad de memoria que requieren. Existen dos enfoques para abordar esta tarea. El primero son las **pruebas de evaluación (benchmarking)** , ejecutar los algoritmos en un computador y medir la velocidad en segundos y el consumo de memoria en bytes. El segundo enfoque depende del **análisis de algoritmos**, independientemente de la implementación y entrada en particular. El primer paso del análisis es abstraer la entrada, para encontrar algún parámetro o parámetros que caractericen el tamaño de la entrada. El segundo paso es abstraer la implementación, para encontrar alguna medida que refleje el tiempo de ejecución del algoritmo, pero que no esté ligado a un compilador o computador en particular. La notación O( ) nos proporciona lo que llamamos un **análisis asintótico.** Sin lugar a dudas, podremos decir que, como n se aproxima asintóticamente al infinito, un algoritmo O( ) es mejor que un algoritmo O(n2). Una sola cifra de prueba de evaluación no podría corroborar dicha afirmación. La notación O( ) abstrae los factores constantes, lo cual facilita su utilización, aunque de forma menos precisa, que la notación T( ). El campo del **análisis de complejidad** analiza problemas, no algoritmos. La primera gran división se realiza entre los problemas que se pueden resolver en tiempo polinomial y los problemas que no pueden resolverse en tiempo polinomial, sin importar el algoritmo que se use. La clase de los problemas polinomiales, los que se pueden resolver en tiempo O(nk) para k, se llama P. Todo el que esté interesado en decidir si P = NP, tendrá que estudiar una subclase de NP llamada problemas **completos NP.** La palabra completos se utiliza en el sentido de los más extremos y se refiere por tanto a los problemas más difíciles de la clase NP. Se ha demostrado que o están en P todos los problemas completos de NP o no está ninguno. La clase **co-NP** es el complemento de NP, en el sentido de que, para todos los problemas de decisiones en NP, existe un problema correspondiente en co-NP con las respuestas si y no invertidas. Sabemos que P es un subconjunto tanto de co-NP como de NP, y se cree que hay problemas en co-NP que no están en P. Los problemas **completos co-NP** son los más difíciles de co-NP.

Vectores, matrices y álgebra lineal
Los matemáticos definen un **vector** como un miembro de un espacio vectorial, sin embargo, nosotros vamos a utilizar una definición mucho más concreta: un vector es una secuencia ordenada de valores. Las dos operaciones fundamentales de los vectores son la suma de vectores y la multiplicación escalar. La longitud de un vector se denota —x— y se calcula tomando la raíz cuadrada de la suma de los cuadrados de los elementos. Los vectores se suelen interpretar como segmentos de líneas dirigidas (flechas) en un espacio n-dimensional. Una **matriz** es un array rectangular de valores organizada en filas y columnas. Esta es la matriz m de tamaño 3 x 4. La suma de las dos matrices se define sumando los elementos correspondientes. También podemos definir la multiplicación de

una matriz por un escalar. Las matrices se utilizan para resolver sistemas de ecuaciones lineales mediante un proceso llamado **eliminación Gauss-Jordan**, un algoritmo O(n3).

Distribuciones de probabilidades
Una probabilidad es una medida sobre un conjunto de sucesos (eventos) que satisface tres axiomas:

- La medida de cada suceso está entre 0 y 1.
- La medida del conjunto completo es 1.
- La probabilidad de una unión de sucesos disjuntos es la suma de las probabilidades de los sucesos individuales.

Un modelo probabilístico consiste en un espacio muestral de resultados posibles mutuamente excluyentes, junto con una medida de probabilidad para cada resultado. **Una función de densidad de probabilidad**, que denotamos también como P(X), pero que tiene un significado ligeramente diferente de la función de probabilidad discreta P(A). También definimos una **función de densidad de probabilidades acumulada** F(X), que es la probabilidad de que una variable aleatoria sea menor que x:

$$F(X) = \int_{-\infty}^{x} P(Z)dz.$$

Una de las distribuciones de probabilidad más importantes es la **distribución gaussiana**, también conocida como la **distribución normal**. Una distribución Gaussiana, con media y desviación estándar (y por tanto una varianza 2) se define como.

$$P(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/(2\sigma^2)},$$

en donde x es una variable continua que va desde - a +. Con 0 medio y discrepancia 2=1, obtenemos el caso especial de la **distribución normal estándar.** Para una distribución sobre un vector x en d dimensiones, existe la **distribución gaussiana multivariada:**

$$P(\mathbf{x}) = \frac{1}{(2\pi)^n|\Sigma|} e^{-\frac{1}{2}((\mathbf{x}-\mu)^\top\Sigma^{-1}(\mathbf{x}-\mu))}$$

en donde es el vector media y es la **matriz de co-varianza** de la distribución.

$$F(X) = \int_{-\infty}^{x} P(x)dx = \frac{1}{2}\left(1 + \text{erf}\left(\frac{x-\mu}{\sigma\sqrt{2}}\right)\right),$$

En una dimensión, también podemos definir la **distribución acumulada** F(x) como la probabilidad de que una variable aleatoria sea menor que x. Para la distribución normal estándar, esto está dado como.

en donde erf(x) es la llamada **función de errores,** que no tiene representación formal cerrada.

El teorema de límite central afirma que la media de n variables aleatorias tiende a una distribución normal a medida que n tiende al infinito. Esto se mantiene para cualquier grupo de variables aleatorias, a menos que la varianza de cualquier subconjunto finito de variables domine a las otras.