**Introducción**

**Fundamentos filosóficos**

Los filósofos definen la hipótesis de la IA débil como la afirmación de que es posible que las máquinas actúen con inteligencia (como si fueran inteligentes); Y la IA fuerte como la afirmación de que las máquinas sí piensan realmente (opuesto al pensamiento simulado).

**26.1 IA débil: ¿pueden las máquinas actuar con inteligencia?**

Algunos filósofos han intentado demostrar que la IA es imposible; que las máquinas no tendrán la posibilidad de actuar inteligentemente. Algunos han utilizado argumentos que tratan de dar el alto a la investigación en IA.

Obviamente, si la IA es imposible o no lo es, dependerá de cómo se defina. Consideremos las siguientes dos cuestiones:

• ¿Pueden volar las máquinas?

• ¿Pueden nadar las máquinas?

La respuesta a la primera pregunta podría ser que, ya que existen aviones y otras máquinas voladoras, más sin embargo no existen maquinas que naden solo barcos que flotan y submarinos que se mueven por debajo del agua, pero esto último ¿no es nadar? , con esto vemos que la respuestas no tienen mucho que ver con el diseño o con las características de los aviones o de los submarinos, y sin embargo sí tienen que ver mucho más con la forma en que se han elegido utilizar las palabras “volar” y “nadar”.

La posibilidad práctica de las “máquinas pensantes” lleva viviendo con nosotros durante sólo 50 años o así, tiempo insuficiente para que los angloparlantes se decidan a dar un significado a la palabra “pensar”.

Alan Turing sugirió que en vez de preguntar si las máquinas pueden pensar, deberíamos preguntar si las máquinas pueden aprobar un test de inteligencia conductiva (test de comportamiento), conocido como el “Test de Turing”.

La prueba del Test de Turing consiste en que el programa mantenga una conversación durante cinco minutos (mediante mensajes escritos en línea, online) con un interrogador (interlocutor). Éste tiene que averiguar si la conversación se está llevando a cabo con un programa o con una persona; si el programa engaña al interlocutor un 30 por ciento del tiempo, este pasará la prueba.

Algunas personas han sido engañadas durante cinco minutos; por ejemplo, el programa ELIZA y el chatbot en Internet llamado MGONZ han engañado a personas ignorantes que no se daban cuenta de que estaban hablando con un programa; el programa ALICE engañó a un juez en la competición del Loebner Prize en el año 2001. Sin embargo, ningún programa se ha acercado al criterio del 30 por ciento frente a jueces con conocimiento, y el campo en su conjunto de la IA no ha prestado mucha atención a los tests de Turing.

Turing examinó una gran gama de posibles objeciones ante la posibilidad de las máquinas inteligentes, incluyendo virtualmente aquellas que han aparecido medio siglo después de que apareciera su artículo algunas de ellas son las siguientes:

**El argumento de incapacidad**

El argumento de incapacidad afirma que una máquina nunca puede ser amable, tener recursos, ser guapo, simpático, tener iniciativas, tener sentido del humor, distinguir lo correcto de lo erróneo, cometer errores, enamorarse, disfrutar con las fresas con nata, hacer que otra persona también se enamore, aprender de la experiencia, utilizar palabras de forma adecuada, ser el tema de su propio pensamiento, tener tanta diversidad de comportamientos como el hombre, hacer algo realmente nuevo.

**La objeción matemática**

Es bien conocido, a través de los trabajos de Turing (1936) y Gödel (1931), que ciertas cuestiones matemáticas, en principio, no pueden ser respondidas por sistemas formales concretos.

En resumen, para cualquier sistema axiomático formal F lo suficientemente potente como para hacer aritmética, es posible construir una

“sentencia Gödel” G(F) con las propiedades siguientes:

• G(F) es una sentencia de F, pero no se puede probar dentro de F.

• Si F es consistente, entonces G(F) es verdadero.

**El argumento de la informalidad**

Consiste en que el comportamiento humano es demasiado complejo para poder captarse mediante un simple juego de reglas y que debido a que los computadores no pueden nada más que seguir un conjunto (juego) de reglas, no pueden generar un comportamiento tan inteligente como el de los hombres.

**26.2 IA fuerte: ¿pueden las máquinas pensar de verdad?**

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.

Esto es lo que Turing llama el argumento de la consciencia, la máquina tiene que ser consciente de sus propias acciones y estados mentales.

Aunque la consciencia sea un tema importante, el punto de vista clave de Jefferson se relaciona realmente con la fenomenología, o el estudio de la experiencia directa, es decir, la máquina tiene que sentir emociones realmente. Otros se centran en la intencionalidad, esto es, en la cuestión de si las creencias, deseos y otras representaciones supuestas de la máquina son de verdad algo que pertenece al mundo real.

Turing reconoce que la cuestión de la conciencia es difícil, pero niega que sea relevante para la práctica de la IA: “ No quiero dar la impresión de que pienso que no hay misterio en torno a la conciencia… Sin embargo no creo que estos misterios tengan necesariamente que resolverse antes de la respuesta a la cuestión que estamos tratando en este trabajo”. Coincidimos con Turing en que nos interesa crear programas que se comporten de forma inteligente y no en si alguien los declara reales o simulados.

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. Desde luego, todavía no hemos dicho lo que significa realmente “isomórficos”, pero la suposición es que existe algún nivel de abstracción por debajo del cual no importa una implementación específica; siempre que los procesos sean isomórficos hasta este nivel, tendrán lugar los mismos estados mentales.

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 (no especificadas) de las neuronas. Así pues, los estados mentales no se pueden duplicar justo en la base de algún programa que tiene la misma estructura funcional con el mismo comportamiento de entrada y salida; necesitaríamos que el programa se ejecutara en una arquitectura con la misma potencia causal que las neuronas. La teoría no dice por qué las neuronas tienen esta potencia causal, ni tampoco qué otras instanciaciones físicas podrían tenerla o no.

Para investigar el punto de vista de estas dos teorías examinaremos uno de los problemas más antiguos de la filosofía la mente, y retomaremos tres experimentos pensados en esta:

**El problema de mente-cuerpo**

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.

René Descartes quien abordó el tema de cómo un alma inmortal interactúa con un cuerpo mortal, y concluyó diciendo que el alma y el cuerpo son dos tipos de cosas diferentes, una **teoría dualista**. **La teoría monista**, frecuentemente llamada **materialismo**, mantiene que no existen cosas tales como almas inmateriales, sino sólo objetos materiales. Como consecuencia, los estados mentales, tales como sentir dolor, saber que alguien está montando a caballo, o creer que la capital de Austria es Viena, son estados del cerebro.

Vamos a cuestionar si es posible que los computadores tengan **estados intencionales** (estos estados son tales como creer, conocer, desear, temer, y otros) que nos ayuden a entender cómo caracterizar dichos estados. Por ejemplo, se podría decir que el estado mental de desear una hamburguesa difiere del estado de desear una pizza porque en el mundo real una hamburguesa y una pizza son cosas diferentes. Es decir, los estados intencionales tienen una conexión necesaria con otros objetos del mundo externo. Por otro lado, anteriormente, hemos argumentado que los estados mentales son estados del cerebro, y de aquí que los estados mentales de identidad o no-identidad se deberían determinar permaneciendo completamente “dentro de la Cabeza”, sin hacer referencia al mundo real.

**El experimento del “cerebro en una cubeta”**

Se imaginan que al nacer alguien le extraen el cerebro de su cuerpo y lo colocan en una cubeta con una ingeniería maravillosa. Esta cubeta mantiene su cerebro y le permite crecer y desarrollarse. Al mismo tiempo, su cerebro recibe unas señales electrónicas de un simulador informático que pertenece a un mundo totalmente ficticio, y las señales motoras de su cerebro son interceptadas y utilizadas para modificar la simulación cuando sea adecuado. A continuación, el estado del cerebro podría tener el estado mental MueroPor (Yo, Hamburguesa) aunque no tenga un cuerpo con el que sentir hambre ni sentido del gusto para experimentarlo, y puede que tampoco haya hamburguesas en el mundo real. En ese caso, ¿sería el mismo estado mental que el del cerebro en un cuerpo?

Una forma de resolver el dilema es decir que el contenido de los estados mentales puede ser interpretado desde dos puntos de vista diferentes. La visión del “**contenido extenso**” interpreta el dilema desde el punto de vista de un observador omnisciente desde fuera con acceso a la situación completa, y que puede distinguir las diferencias del mundo. De esta manera, bajo el “**contenido extenso”**, las ideas del cerebro en una cubeta son diferentes de las de una persona normak. El contenido estrecho sólo tiene en cuenta el punto de vista subjetivo interno, y bajo este punto de vista todas las creencias serían las mismas.

**El experimento de la prótesis cerebral**

Supongamos que una técnica quirúrgica milagrosa puede sustituir las neuronas individuales con los mecanismos electrónicos sin interrumpir el funcionamiento del cerebro por completo. El experimento consiste en sustituir gradualmente todas las neuronas de la cabeza de alguien con mecanismos electrónicos y a continuación invertir el proceso para retornar al sujeto a su estado biológico normal.

Nos preocupa tanto el comportamiento externo como la experiencia interna del sujeto, durante y después de la operación. Por definición del experimento, el comportamiento externo del sujeto no debe sufrir ningún cambio en comparación con lo que se observaría si la operación no se llevase a cabo. Ahora bien, aunque la presencia o ausencia de consciencia no la pueda asegurar fácilmente un tercero, el sujeto del experimento debería por lo menos poder registrar cualquier cambio en su propia experiencia consciente.

Moravec, un investigador y funcionalista en robótica, está convencido de que su consciencia no se vería afectada. Searle, un filósofo y naturalista biólogo, está convencido de que la consciencia desaparecería:

Para que el comportamiento externo siga siendo el mismo, mientras que el individuo se va sumiendo gradualmente en la inconsciencia, debe darse el caso de que la propia voluntad del sujeto se elimine instantáneamente y totalmente; de otra manera, la reducción del conocimiento se vería reflejada en el comportamiento externo, con palabras tales como “Ayuda, mi conocimiento se está reduciendo”, u otras palabras con ese mismo efecto. Esta eliminación instantánea de la propia voluntad como resultado de la sustitución gradual de neuronas una a una parece ser una afirmación improbable.

Por las condiciones del experimento, obtendremos respuestas del tipo, “Me encuentro bien. Debo decir que estoy algo sorprendido porque creía en el argumento de Searle”. O de lo contrario, podríamos atizar al sujeto con un bastón acabado en punta y observaríamos la respuesta, “¡Ay, cómo duele!”. Ahora bien, durante el curso normal de sucesos, el escéptico puede desechar tales salidas de programas de IA tales como simples artimañas. Ciertamente, es bastante fácil utilizar una regla tal como: Si el sensor 12 identifica “Alto” entonces la salida será “Ay”. Sin embargo, lo importante es que, debido a que hemos duplicado las propiedades funcionales de un cerebro humano normal, supondremos que el cerebro electrónico no contiene dichas artimañas.

El comportamiento externo del sujeto debe ser, por definición, como si la operación no hubiera ocurrido. En particular, deberíamos poder preguntar, “¿Cómo estaba durante la operación?, ¿recuerda el bastón acabado en punta?”. El sujeto debe tener recuerdos precisos de la naturaleza real de sus experiencias conscientes a pesar del hecho de que según Searle no ha habido tales experiencias.

Searle podría responder que no hemos definido el experimento adecuadamente. Por ejemplo, si las neuronas reales quedan suspendidas entre el momento que se extraen y el tiempo en que se reemplazan en el cerebro, por supuesto entonces no “recordarán” las experiencias durante la operación. Para tratar esta eventualidad, necesitamos asegurarnos de que el estado de las neuronas se actualiza para reflejar el estado interno de las neuronas artificiales que están reemplazando. Si los supuestos aspectos no funcionales de las neuronas reales dan como resultado entonces un comportamiento funcionalmente diferente del que se observa con las neuronas artificiales todavía en su lugar, tenemos un reductio ad absurdum simple, debido a que eso significaría que las neuronas artificiales no son equivalentes funcionalmente a las neuronas reales.

**La habitación china**

El sistema se compone de un hombre, que solamente entiende inglés, y que está equipado con un libro de reglas escrito en inglés y varias pilas de papel, algunas en blanco y algunas con inscripciones indescifrables. (El hombre entonces hace el papel de la CPU, el libro de normas es el programa y los papeles son el dispositivo de almacenamiento.)

El sistema se encuentra dentro de una habitación con una apertura al exterior. A través de la apertura se van deslizando los papeles con símbolos indescifrables. El hombre encuentra los símbolos correspondientes en el libro de reglas y sigue las instrucciones. Las instrucciones pueden incluir escritura de símbolos en los papeles nuevos que van saliendo, encontrar símbolos en las pilas de papeles, reorganizar las pilas, etc. Finalmente, las instrucciones harán que un símbolo o más sean transcritos a un trozo de papel que se pasa otra vez al mundo exterior.

Hasta ahora todo está bien. Pero Desde fuera, se observa un sistema que está sacando la entrada en forma de sentencias chinas y generando respuestas chinas que obviamente son tan “inteligentes” como las de la conversación imaginada por Turing.

Entonces Searle argumenta lo siguiente: la persona que está en la habitación no entiende el chino (supuestamente). El libro de reglas y las pilas de papel, que son sólo trozos de papel no entienden el chino. Por consiguiente, no está habiendo comprensión del chino. De aquí que, según dice Searle, ejecutar el programa adecuado no genera necesariamente entendimiento.

La objeción es que aunque se pueda preguntar si el hombre de la habitación entiende el chino, esto es como preguntar si la CPU puede admitir raíces cúbicas. En ambos casos, la respuesta es no, y en ambos casos según la respuesta del sistema, el sistema completo sí tiene la capacidad en cuestión.

Reiterando, el objetivo del argumento de la habitación china es refutar la IA fuerte, es decir la afirmación de que ejecutar la clase adecuada de programa da como resultado necesariamente una mente. Esto se hace exhibiendo un sistema aparentemente inteligente que ejecuta la clase adecuada de programa que no es, según Searle, demostrablemente una mente.

Searle permite la posibilidad lógica de que el cerebro esté implementando de verdad un programa de IA de la clase tradicional. Sin embargo, si el mismo programa se ejecuta en la clase inadecuada de máquina no sería una mente. Searle no cree que las “máquinas no puedan tener mentes”, en cambio afirma que algunas máquinas sí que tienen mentes**, los hombres son máquinas biológicas con mentes.**

**26.3 La ética y los riesgos de desarrollar la Inteligencia Artificial**

Todos los científicos e ingenieros se enfrentan a consideraciones éticas de cómo deberían actuar en el trabajo, qué proyectos deberían o no deberían hacer, cómo los deberían abordar. Sin embargo, la IA parece que expone nuevos problemas como:

• 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.

La IA tiene su contraparte con respecto a estos posibles problemas, por ejemplo:

* Hasta ahora, la automatización por medio de la tecnología de la IA ha creado más trabajos de los que ha eliminado, y ha creado puestos de trabajo más interesantes y mejor pagados.
* La IA incrementa el ritmo de la innovación tecnológica y contribuye así a esta tendencia general, pero la IA también mantiene la promesa de permitirnos ahorrar tiempo y permitir que nuestros agentes automatizados hagan las cosas por un tiempo.
* La IA, aunque sea una materia de gran éxito, quizá sea por lo menos amenazante para las suposiciones morales de la sociedad del siglo XXI al igual que la teoría de la evolución lo fue para los del siglo XIX.
* Si los robots adquieren consciencia, tratarlos entonces como meras “máquinas” (por ejemplo, tratarlos como algo aparte) podría ser inmoral. Los robots también deben actuar moralmente, necesitaríamos programarlos con una teoría de lo que está bien y lo que está mal.

IA: presente y futuro

**27.1 Componentes de los agentes**

Se examinará el agente basado en la utilidad ya que este es el diseño de agentes más general de todos. Se estudiará también su extensión con capacidades de aprendizaje.

**Interacción con el entorno a través de sensores y actuadores**

Los sistemas IA se construyeron de tal forma que los humanos tenían que proporcionar las entradas e interpretar las salidas, mientras que los sistemas robóticos se centraban en tareas de bajo nivel en donde el razonamiento y la planificación de alto nivel estaban totalmente ausentes. Esto fue debido en parte al gran esfuerzo que requería la ingeniería y el alto coste para hacer funcionar robots listos.

En los últimos años la situación ha cambiado rápidamente con la disponibilidad de robots programables listos para utilizarse o prefabricados. A su vez, estos se han beneficiado de las pequeñas cámaras CCD baratas de alta resolución, y de las unidades motoras compactas y fiables.

**Seguir la pista del estado del mundo**

Esta es una de las capacidades centrales que se requieren para un agente inteligente. Requiere tanto percepción como actualización de las representaciones internas.

Los algoritmos actuales de filtrado y de percepción pueden combinarse para hacer un trabajo razonable de informar predicados de bajo nivel, tales como “la taza está sobre la mesa”, sin embargo, tenemos mucho que recorrer antes de que puedan informar de que “el Doctor Russell está tomando una taza de té con el Doctor Norvig”.

**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. Algunos de los algoritmos de planificación utilizan representaciones jerárquicas y representaciones de primer orden para manejar problemas de esta escala.

**La utilidad como expresión de preferencias**

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. Sin embargo, hasta ahora, no se ha trabajado mucho en la construcción de funciones de utilidades realistas; imaginemos, por ejemplo, una red compleja de preferencias interactuantes que debe ser entendida por un agente que funciona como un asistente de oficina para un ser humano. Resulta muy difícil descomponer las preferencias en estados complejos de la misma forma que las redes de Bayes descomponen las creencias sobre los estados complejos; Hay que tomar muy en serio la tarea de ingeniería del conocimiento para las funciones de recompensa como una forma de transmitir a nuestros agentes qué es lo que queremos que hagan.

**Aprendizaje**

Se han desarrollado técnicas estadísticas y lógicas muy poderosas que pueden hacer frente a problemas bastante grandes, alcanzando o excediendo las capacidades humanas en la identificación de patrones predictivos definidos en un vocabulario dado. Por otro lado, el aprendizaje de la máquina no ha avanzado mucho en cuanto al importante problema de construir representaciones nuevas a niveles de abstracción mayores que el vocabulario de entrada.

**27.2 Arquitecturas de agentes**

Hemos visto que para situaciones en las que el tiempo no es esencial, se necesitan respuestas reflejas, mientras que la deliberación basada en el conocimiento permite que el agente planifique con antelación. 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.

Los agentes también necesitan formas de controlar sus propias deliberaciones. Deben poder parar de deliberar cuando se exige la acción, y deben poder utilizar el tiempo del que disponen para la deliberación y ejecutar los cómputos más rentables. Por ejemplo, un agente que conduce un taxi y ve un accidente delante, debe decidir en menos de un segundo si frenar o actuar de forma evasiva. También debería pensar en menos de un segundo las cuestiones más importantes, tales como si los carriles de la izquierda y derecha están libres y si tiene un camión detrás muy cerca, en vez de preocuparse de la tensión y el desgaste en las ruedas (neumáticos, gomas) o de dónde recoger al siguiente pasajero.

En los últimos años han surgido dos técnicas prometedoras. La primera conlleva la utilización de algoritmos de cualquier momento (anytime), donde 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 (en lo que se refiere a demorar la acción) como de sus beneficios (en cuanto a la calidad de decisión mejorada).

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.

**27.3 ¿Estamos llevando la dirección adecuada?**

El objetivo de la IA es que se quiere construir agentes, pero, ¿qué especificación se debe tener en mente? Existen cuatro posibilidades:

**Racionalidad perfecta**

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 calculadora**

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. Esta es una propiedad interesante para que un sistema la presente; sin embargo en la mayoría de los entornos, la respuesta adecuada no tiene valor en el momento equivocado

**Racionalidad limitada**

La racionalidad limitada funciona en primer lugar satisfaciendo, es decir, deliberando sólo el tiempo necesario para elaborar una respuesta que sea lo “suficientemente buena”. En muchos casos, parece que es un modelo útil para los comportamientos humanos. Sin embargo, no es una especificación formal para los agentes inteligentes, porque la definición de “suficientemente bueno” no la ofrece esta teoría. Además, satisfacer parece ser sólo uno de la gran cantidad de métodos que se utilizan para hacer frente a recursos limitados.

**Optimalidad limitada**

se propone un concepto de optimalidad limitada como una tarea formal para la investigación de la IA que esté bien definida y que sea viable. La optimalidad limitada especifica programas óptimos en vez de acciones óptimas. Las acciones, después de todo, son generadas por programas, y es sobre los programas en donde tienen control los diseñadores.

**27.4 ¿Qué ocurriría si la IA tuviera éxito?**

La IA ha hecho posibles aplicaciones nuevas tales como los sistemas de reconocimiento de voz, sistemas de control de inventarios, sistemas de vigilancia, robots y motores de búsqueda. Se puede esperar que los éxitos de medio nivel en IA afectarán a toda clase de gente en sus vidas cotidianas. Hasta ahora, las redes de comunicaciones computarizadas, tales como los teléfonos móviles e Internet, han tenido este tipo de efecto penetrante o dominante en la sociedad, pero la IA no.

Finalmente, parece probable que un éxito en IA a gran escala, la creación de inteligencia en el nivel humano y más allá, cambiaría las vidas a una mayoría de la humanidad. La verdadera naturaleza de nuestro trabajo y de nuestro papel cambiaría, así como nuestro punto de vista de la inteligencia, la consciencia y el destino futuro de la raza humana. A este nivel, los sistemas IA podrían suponer una amenaza directa a la autonomía humana, la libertad, e incluso la supervivencia. Por estas razones, la investigación en IA no se puede divorciar de sus consecuencias éticas.