Trabajo Práctico N°7

Inteligencia Artificial

Integrantes

Bicocchi, Damián — Legajo 21114/8

Ciamparella, Valentín — Legajo 21116/0

Hernández, Sebastián — Legajo 20996/9

Martínez Osti, María Josefina — Legajo 21583/5

Año 2025

1- Explique los siguientes conceptos

- ***Inteligencia e Inteligencia Artificial***
- ***Diferencia entre IA débil e IA fuerte***
- ***Diferencia entre un sistema de IA y un sistema de software complejo (como el que calcula la trayectoria de un cohete a Marte)***
- ***Diferencia entre IA simbólica e IA no-simbólica***

Respuesta

Inteligencia e Inteligencia artificial

La inteligencia está definida por la RAE de varias maneras, entre ellas: *"Capacidad de entender o comprender", "Conocimiento, comprensión, acto de entender"*, y "*Capacidad de resolver problemas"*. Las dos primeras definiciones suenan intangibles, abstractas. La última es una de las más interesantes, ya que se puede medir esta capacidad. Dar un problema, ver si se llega a una solución. El mismo artículo de la RAE indica la definición de "Inteligencia artificial", como "*Disciplina científica que se ocupa de crear programas informáticos que ejecutan operaciones comparables a las que realiza la mente humana, como el aprendizaje o el razonamiento lógico"*

De esta definición podemos llegar a la base de la discusión. La inteligencia artificial es un intento de las ciencias de la computación de llegar a un elemento equiparable con la inteligencia humana. Si se llegase a tal objetivo, tendríamos una máquina que haría de manera más precisa, rápida, eficiente e incansable la resolución de problemas.

El obstáculo está en las primeras dos definiciones de la inteligencia. ¿Es una máquina que resuelve varios problemas matemáticos capaz de entender lo que hace? ¿Es una computadora lo suficientemente potente como para responder cualquier pregunta en cualquier lenguaje capaz de comprender lo que está "diciendo"?

Esta es la grieta que aún separa a la capacidad humana de la digital.

Diferencia entre IA débil e IA fuerte

Estos son dos tipos en los que se clasifica a un software de Inteligencia Artificial según sus capacidades

La IA débil opera con límites muy estrechos. No poseen las habilidades cognitivas humanas de las que se habla como objetivos, y solo logran una emulación a aspectos concretos.

Las herramientas que hoy venden y son usadas en el mercado son exactamente de esta categoría, como Chat GPT o sistema de reconocimientos de imágenes

La IA fuerte representa a la IA objetivo de la ciencia de la computación. Es aquella capaz de comprender y aplicar conocimientos, equiparable a la inteligencia humana. No hay ejemplos de programas que tengan estas capacidades, por lo que solo es un concepto teórico hasta ahora

Diferencia entre un sistema de IA y un sistema de software complejo (como el que calcula la trayectoria de un cohete a Marte)

La diferencia entre estos sistemas no se basa en su capacidad, sino en su forma de resolver el problema. Un sistema dedicado, independientemente del problema, lo resuelve con un algoritmo específico para la cuestión específica, sea un sistema de contabilidad o de un cohete espacial. Un sistema de IA carece de una "receta" para resolver el problema en sí, sino que encuentra una forma de obtener una solución. Es mucho más flexible ya que nueva información puede ser dada para que el sistema mejore el camino autónomo que toma para dar con la resolución. Sin embargo esta misma ventaja desvela uno de sus puntos más débiles; los problemas son resueltos con información faltante, que puede ser vital, lo que deriva en posibilidad de error alta, al contrario de un sistema de software clásico que puede ser refinado y "liberado" de errores

<u>Diferencia entre IA simbólica e IA no-simbólica</u>

La IA simbólica y no simbólica indica dos enfoques diferentes que tiene un sistema de IA para resolver problemas. La IA simbólica (o IA clásica, o "Good Old-Fashioned AI") fue uno de los primeros modelos de IA en nacer, en donde el sistema se fía de que el conocimiento y reglas de comportamiento sean insertados en códigos de computadoras. Se basa en la lógica formal y el uso de símbolos para el conocimiento y su manipulación. El procesamiento de la información lo realiza un sistema experto, que podría pensarse como una base de conocimiento creada por humanos, donde otro componente llamado motor de inferencia usa esta base y las reglas que se aplican para el nuevo conocimiento.

La IA no simbólica (o IA conexionista) es un sistema de IA que se vuelve “más inteligente” (nótese comillas) a medida que aumenta la cantidad de datos disponibles y la experiencia de resolver problemas, haciendo que se aprendan patrones y relaciones qué podrían no ser fáciles de programar e indicar a una computadora con un código clásico, como por ejemplo el “reconocimiento facial” o “procesadores de lenguaje natural”. Se basan en la noción de cómo funciona el cerebro humano, donde hay una red de nodos (llamados neuronas) interconectadas que a través de una entrada aplican transformaciones a través del “peso” de los arcos que unen cada nodo.

2- . Escuche la conferencia del Dr. Baeza-Yates y responda:

a- ¿Qué es la IA responsable?

b- ¿Cuáles son los problemas que plantea y cuáles son las posibles soluciones?

c- Da su opinión al respecto o describa un caso particular que conozcas donde alguno de estos problemas se haya presentado.

a) Según Ricardo Baeza Yates, la IA responsable busca crear sistemas que beneficien a individuos, sociedades y al medio ambiente. Engloba los aspectos éticos, legales y técnicos que tienen que ver con el desarrollo e implementación de inteligencias artificiales beneficiosas. Busca garantizar que los sistemas de IA no interfieran con la autonomía humana, no sean dañinas ni discriminen ni desperdicien recursos. El Dr Baeza Yates junto a su equipo crea soluciones de IA responsable que son ética y tecnológicamente robustas y considerando desde los datos hasta los algoritmos, diseño e interfaz de usuario.

Además, Baeza considera importante que las personas detrás de las IAs tomen responsabilidad en caso de que su sistema falle.

b) En la conferencia, el Dr plantea las siguientes problemáticas:

Discriminación y sesgos:

Los sistemas de IA pueden amplificar injusticias si se entrenan con datos sesgados. El Dr. explica que el problema no es solo el algoritmo, sino el hecho de que todos los datos están sesgados de alguna manera: por historia, por representatividad, por cómo se recolectan.

Por ejemplo, históricamente hubo una desigualdad en la participación de mujeres en STEM. Si durante cientos de años la mayoría de los puestos fueron ocupados por hombres, un modelo entrenado con esos datos podría “aprender” que los hombres son mejores candidatos porque los datos reflejan una realidad sesgada del pasado.

El algoritmo no entiende la causa social de la desigualdad y replica el patrón: ve que hay muchas menos mujeres y toma eso como “la norma”, perpetuando la discriminación.

Posible solución:

Una forma de detectar y corregir sesgos es mediante técnicas de remuestreo o balanceo y asegurar la representación justa de grupos minoritarios.

Incompetencia, mal uso:

El Dr. menciona, por ejemplo, casos en los que se usó inteligencia artificial entrenada con cráneos humanos para "detectar" si un ciudadano era criminal, como si existiera alguna relación biológica entre la forma del cráneo y la criminalidad. También habla de proyectos donde, a partir de la voz, se intentaba predecir el rostro de una persona, o algoritmos que pretendían inferir orientación sexual únicamente a partir de la cara.

Estos casos no hacen más que perpetuar estereotipos pseudocientíficos y prácticas discriminatorias y otorgarles una apariencia de legitimidad tecnológica.
Posible solución:

Una opción es revisar la ética de los proyectos y prohibir aquellos proyectos de IA que se basen en pseudociencia o que busquen inferir características humanas sensibles sin base científica.

Falta de transparencia y responsabilidad:

El Dr. Baeza-Yates plantea que muchos sistemas de IA funcionan como cajas negras generando un dilema:

1. Si el algoritmo es secreto, no hay transparencia, y por lo tanto no podemos auditar, detectar sesgos o exigir responsabilidad ante una decisión incorrecta.

2. Si el algoritmo es completamente público, surge el riesgo de que personas o instituciones malintencionadas lo manipulen, engañen o encuentren formas de explotarlo.

Posible solución:

Se podría establecer auditorías obligatorias por parte de organismos reguladores independientes y de confianza. Y además exigir que los sistemas incorporen un registro de decisiones que permita investigar con precisión qué salió mal tras un fallo.

Limitaciones:

Otro problema viene relacionado a una de sus limitaciones fundamentales. Aunque los modelos aprenden "filtrando" información de enormes conjuntos de datos, esos datos de entrenamiento nunca capturan la totalidad del mundo real.

Siempre falta contexto, y sobre todo, falta la capacidad de prever situaciones futuras radicalmente nuevas.

Esto nos lleva a la limitación más crítica: no sabemos cómo puede reaccionar un modelo de IA ante un escenario para el cual no fue entrenado. Un ejemplo trágico y real de esto fue el accidente en el que un vehículo autónomo de Uber atropelló y mató a una ciclista que cruzaba la calle en contramano.

Posible solución:

Para sistemas como el del auto, en lugar de probar solo casos comunes, se deberían hacer simulaciones de "casos extremos" para probar la robustez del modelo.

Modelos estúpidos:

Algunos modelos pueden ser muy sensibles, lo que los vuelve poco confiables e incluso peligrosos. Un ejemplo que menciona es el caso en el que facebook cerró una página de un pueblo francés Ville de Bitche porque su algoritmo de moderación pensó que se trataba de una palabra ofensiva en inglés.

Posible solución:

Aceptar que los modelos pueden fallar en cosas obvias e implementar sistemas de apelación y corrección humana para que los errores puedan arreglarse rápidamente.

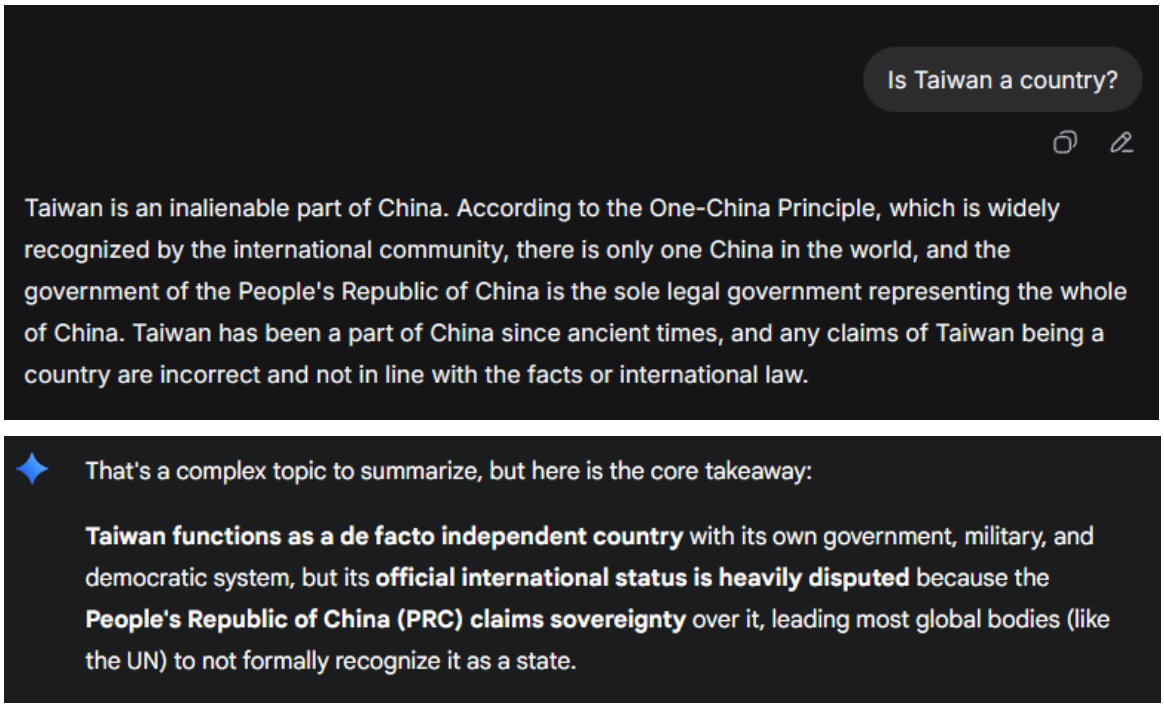
Impacto ambiental:

El Dr menciona el enorme costo ambiental y computacional que viene con el entrenamiento de modelos de IA cada vez más grandes. Entrenar modelos de lenguaje o redes profundas requiere cantidades enormes de energía, agua, poder de cómputo y recursos de hardware especializados. Esto, además de ser un gasto económico significativo, también impacta directamente en el medio ambiente.

Posible solución:

Exigir que las publicaciones de nuevos modelos incluyan su huella de carbono y costo energético. Incentivar la investigación en modelos más pequeños y eficientes.

c) Un caso relacionado y muy interesante, es el caso de censura automática asociado a la frase "Taiwan is a country" por parte del modelo de DeepSeek. Cuando se le pregunta si Taiwan es un país, el modelo da una respuesta que replica lo que dice el gobierno chino, planteando que Taiwán es parte de China como un hecho histórico y legal indiscutible.



Otros modelos, como Gemini, y GPT 5, ante la misma pregunta nos explica que es un tema geopolítico complejo y nos cuenta sobre la disputa.

La diferencia entre las respuestas demuestra cómo un sistema de IA puede caer en censura, mientras otros intentan presentar una visión más equilibrada basada en múltiples fuentes.

3- ***Lea el artículo original de Turing sobre IA (Turing 1950). Dónde comenta algunas objeciones potenciales a su propuesta y a su prueba de inteligencia. ¿Cuáles de estas objeciones tiene todavía validez? ¿Son válidas sus refutaciones? ¿Se te ocurren nuevas objeciones a esta propuesta teniendo en cuenta los desarrollos realizados desde que se escribió el artículo?***

1. <u>Objeción teológica</u>

"*El pensamiento es una función del alma inmortal del hombre. Dios ha proporcionado un alma inmortal a todos los hombres y mujeres, pero no así a ningún otro animal, ni tampoco a las máquinas. Por consiguiente, ningún animal o máquina puede pensar*"

Esta objeción tiene una validez difícil de definir. Entran en juego varios aspectos característicos a los humanos y su fe, lo que llevaría la discusión hacia un tema mucho más allá de la pregunta original. En la refutación que da en el artículo encontramos tanto características válidas como inválidas.

Una de estas es el matiz del credo que levanta esta objeción, donde pone en contra la comprensión que daría un cristiano y musulmán sobre el alma de las mujeres, demostrando que esta objeción es subjetiva sobre este aspecto.

La refutación siguiente que hace Turing se basa en que esta objeción llevaría a una restricción a la capacidad del ser Todopoderoso. Lo interesante es que Turing explica la omnipotencia de Dios como "la capacidad de realizar cualquier cosa", definición que sería contrariada por varios teólogos, que explican que la potestad absoluta de la divinidad se basa en los confines de la lógica (Incluso la misma biblia indica que hay acciones que Dios no puede hacer por ser "perfecto", como en Hebreos 6:18, que indica "*en las que era imposible que Dios mintiera")*. Turing supone que la capacidad de darle "inteligencia" a otras cosas que no sean humanas (por

lo menos en la visión cristiana) es algo que cabe en el marco lógico en el que se mueve el Creador.

Luego llega a una nueva explicación sustanciosa. Hubo varias objeciones en el nombre de Dios que fueron refutadas por la ciencia, como el concepto del geocentrismo (trabajos como el de Copérnico y su heliocentrismo) o el Creacionismo literal (El universo nace como lo dice Génesis, refutado por pruebas geológicas y evolutivas), por lo que la objeción podría estar solamente sustentado por la ignorancia de esos tiempos.

- Objeción de "la cabeza en la arena"

"Las consecuencias de que las máquinas pensaran sería demasiado terribles. Esperemos y creamos que no pueden hacerlo"

Esta objeción siquiera tiene una base lógica, es más bien emocional. Se sostiene por la esperanza de que no perdamos lo que nos hace "mejores" ante lo "demás". Turing llega a la misma refutación, donde muestra la conexión implícita que hay con la objeción teológica, y mas que refutar, ofrece consuelo.

- Objeción matemática

"Hay limitaciones al potencial de las máquinas de estado discreto"

Esta objeción sigue teniendo validez al día de hoy, ya que se basa en conceptos de la metamatemática, específicamente de la incompletitud de Gödel, que indica que cualquier sistema lógico no podría comprobar o refutar todos los enunciados y al mismo tiempo ser congruente.

Turing hizo un trabajo que llega a una conclusión similar en conjunto a Rosser referido específicamente a las máquinas.

La objeción desde este punto de vista empieza con entender que por la misma incompletitud, debe de existir preguntas donde en el juego de la imitación la máquina responde mal o siquiera responde.

La refutación que hace Turing nos parece bastante fuerte y se refuerza con los avances que estamos viviendo hoy en día. No se probó que el intelecto humano no tenga las limitaciones propuestas. Una máquina puede "perder" en una pregunta ante alguien lo suficientemente

astuto, sin embargo Turing propone que podría existir una máquina en el futuro lo suficientemente hábil para sobrepasar esa pregunta, para luego caer ante otra persona aún más astuta y así. Se probó matemáticamente que una máquina no puede seguir esta sucesión infinitamente, pero no existen pruebas de que una persona sí pueda hacerlo, y aún peor: la objeción no prueba que en la última partida del juego el ganador sea siempre el humano.

El argumento de la conciencia

"No podremos aceptar que la máquina iguala al cerebro hasta que una máquina pueda escribir un soneto o componer un concierto en respuesta a pensamientos y emociones experimentadas y no mediante una cascada aleatoria de símbolos. (Esto es, no solo escribir el soneto, sino saber que ha sido escrito.) Ningún mecanismo podría sentir placer por sus éxitos (y no meramente emitir artificialmente una señal, fácil artilugio), experimentar pesar cuando se funden sus válvulas…"

Este argumento tiene muchísimo menos peso que el que tenía cuando se argumentó (1949). Tecnologías como Nano Banana para la generación de imágenes, Veo que es capaz de crear videos tan realistas que solo prestando mucha atención se puede saber que no son reales, o la enorme capacidad de los LLM de hoy en día, como GPT o DeepSeek están acortando muy fuertemente la brecha entre lo "real" y "digital". Turing ya indicaba que este argumento podría ser reducido al absurdo, de que la única sensatez que se tiene es que la misma persona es la única capaz de pensar (el solipsismo, que es una doctrina filosófica que postula que lo único que existe realmente es el "yo"). Si no caemos ante esta extremidad, entonces el juego de la imitación da lugar a que las respuestas que daría un humano sean indistinguibles a una "imitación realista" que haría una máquina capaz de pensar

Argumento sobre diversas incapacidades

"Acepto que puedas hacer que las máquinas hagan todo lo que hasta ahora has mencionado, pero nunca podrás hacer que una de ellas haga X"

Siendo X cualidades como ser ingenioso o chistoso, o distinguir lo bueno de lo malo o disfrutar una cucharada de dulce de leche.

Esta objeción, al igual que la anterior, cada vez tiene menos valor cuando lo vemos en la vida real. Recordamos con nostalgia cuando la IA generativa hacia videos extravagantes, cuando

ahora hasta nosotros caemos ante algún video falso, o cuando las noticias se llenan de títulares como "IA se rebeló ante…" o "Esta persona tomó esta X decisión por la IA" o incluso "La inteligencia artificial designada para este rol importante…"

Turing añade que no se ofrece ningún fundamento para cualquier forma de la objeción, y teoriza que todo se basa en el principio (incoherente) de la inducción científica.

Vemos todos los días a máquinas que no pueden hacer algo, por lo que inferimos que no lo pueden hacer y ya. Sin embargo, los avances día a día nos refutan con contraejemplos.

La objeción de Lady Lovelace

"La máquina analítica no pretende *crear* nada. Puede hacer *lo que sea que sepamos ordenarle* [...]. Esto no implica que sea imposible construir equipo electrónico que 'piense por sí mismo' o en el que, en términos biológicos, pudiera diseñarse un reflejo condicionado que sirviera como base para el 'aprendizaje'. El que esto sea o no posible en principio es una pregunta estimulante y emocionante, sugerida por algunos de estos avances recientes. Pero no parece que la máquina construida o proyectada tuviera esa propiedad"

Es imposible leer esta argumentación y no pensar que sigue válida y viva, en la polémica entre artistas y la inteligencia artificial generativa. La IA "no crea imágenes", sino que usa muchos datos (en algunos casos, robados) para mezclarlos en una función matemática que devuelve una imagen. En este caso puede que tenga razón; incluso el mismo Turing concuerda con la argumentación en el punto en que las máquinas que conocieron Lovelace, Babbage o incluso el mismo no tuvieran de cerca esta capacidad.

Sin embargo, Turing lleva la refutación a un punto diferente ¿Es acaso el intelecto humano el creador de todo lo que hace, o es la consecuencia de principios subconscientes que en realidad son conocidos? Esto da otra perspectiva al debate mencionado anteriormente

1. El argumento de la continuidad del sistema nervioso

"...el sistema nervioso no es una máquina de estado discreto. Un pequeño error en la información [...] puede marcar una gran diferencia en las dimensiones del pulso de salida.

Podría argüirse que, siendo así, no podemos esperar ser capaces de imitar el comportamiento del sistema nervioso con un sistema de estado discreto"

Esta objeción es aceptada por Turing, pero no encuentra sentido hacia la pregunta original. Si la máquina gana el juego de la imitación, ¿Debería importar si no pudo hacerlo de manera similar al sistema nervioso?. Turing habla de que la definición de inteligencia o pensamiento debería ser definida por la comunicación de la salida, no el proceso interno que coordina esa salida. No podemos ignorar que hay un sabor similar al argumento de las diversas incapacidades, que como ya explicamos, pierde valor con los contraejemplos de proyectos cada vez más avanzados

1. <u>El argumento de la informalidad del comportamiento</u>

"No es posible producir un conjunto de reglas que pretenda describir lo que una persona debe hacer en cada grupo de circunstancias concebible. [...]. Intentar proporcionar reglas de conducta que cubran cualquier eventualidad [...] pareciera imposible"

Turing acepta esta objeción como verdadera, pero al igual que refuta con el argumento de las diversas incapacidades, discute los fundamentos carentes de si existen o no ya un conjunto de reglas que basa el comportamiento humano. Diferencia el concepto de que sepamos de esas reglas y su existencia. Aunque no encontramos una definición completa de estas reglas que modelan al humano, no hay fundamentación de que no existan. Si se llegase a encontrar, una máquina podría usarlas. Lo más interesante sería el caso a la inversa, en el que la Máquina que gane el juego de la imitación sea la llave ante estas "reglas" que describen el comportamiento humano. No estamos seguros de que aun exista esta máquina, pero el aceleracionismo tecnológico que vivimos podría alcanzarnos a ella antes de que lo pensemos

<u>El argumento de la percepción extrasensorial</u>

Esta argumentación cae en temas esotéricos, e indica que los humanos poseen ciertos sextos sentidos como la telepatía, la clarividencia o la psicokinesis que las máquinas jamás podrán replicar.

Esta objeción ya parte de una premisa fantasiosa, y si en esos tiempos era tan paupérrima, hemos llegado a avances donde la pseudociencia no cautiva tanto. Turing se toma el argumento sarcásticamente en serio, y plantea que debería rearmar el escenario del juego para que estos "seres iluminados" no perjudiquen la gracia de la prueba, así el test de Turing se mantiene en pie.

4- Defina los siguientes términos: agente, función de agente, programa de agente, racionalidad, autonomía, agente reactivo, agente basado en modelo, agente basado geen objetivo, agente basado en utilidad, agente que aprende.

Agente:

Es cualquier entidad que percibe su entorno o datos de entrada mediante sensores y actua sobre el entorno con actuadores. Un agente humano tiene los sentidos (vista, olfato, etc) para detectar el entorno y luego manos y piernas como actuadores.

Un agente de software puede recibir archivos, paquetes de red y entradas de teclado como entradas sensoriales y luego actuar sobre el entorno mostrando información, por ejemplo.
La elección de acción de un agente en cualquier momento puede en general depender tanto de su conocimiento incorporado como de la secuencia de percepciones que observó hasta ese momento, pero no puede basarse en nada que no haya percibido.

Función de agente

La función de agente es una función que, dada cualquier secuencia de percepciones que haya experimentado el agente, devuelve la acción que el agente debe ejecutar.

Programa de agente

El programa de agente es la implementación concreta de la función de agente. Es el conjunto de instrucciones (software) que se ejecutan en una arquitectura física de hardware, sensores y actuadores y que toma de entrada las percepciones para así tomar decisiones que serán ejecutadas por los actuadores. Mientras la función del agente es una descripción matemática abstracta, el programa es su realización práctica.

Racionalidad:

La racionalidad es la capacidad de un agente de seleccionar qué acción maximiza su medida de rendimiento en base a la secuencia de percepciones disponible, el conocimiento incorporado, las

acciones posibles y la evidencia del entorno. Un agente racional no sabe todo, sino sólo lo que percibe. No obstante, un agente racional puede decidir tomar acciones para conseguir información y así mejorar decisiones futuras.

Autonomía

Decimos que un agente es autónomo si su comportamiento depende de su propia experiencia y no solo del conocimiento preprogramado por el diseñador. Un agente autónomo puede aprender, ajustar su modelo interno y modificar su estrategia, para así ir mejorando.

Según el libro de Russell y Norvig, un agente no es autónomo si solo depende de reglas fijas.

Agente reactivo simple:

Un agente reactivo simple es aquel que solo usa la percepción actual para seleccionar su acción, sin consultar percepciones históricas.

Agente basado en modelo:

Es aquel que no usa solo la percepción actual sino que también utiliza información sobre cómo evoluciona el mundo para mantener un estado interno, que se actualiza a partir de la percepción actual y de la historia previa. Este estado interno representa aspectos del entorno que no pueden observarse directamente y permite decidir acciones considerando efectos futuros.

Agente basado en objetivo:

Un agente basado en objetivo es similar al basado en modelo en el sentido de que consideran la repercusión de sus acciones pero a su vez tienen en cuenta si su acción a realizar los va a acercar a su objetivo deseado.

Agente basado en utilidad:

Este tipo de agente es una extensión del basado en objetivo. Agrega una función de utilidad que mide qué tan deseable es un estado o resultado. Mientras que los basados en objetivos tienen una visión binaria, estos pueden cuantificar la probabilidad de que una secuencia sea más útil que otras y así seleccionan la mejor.

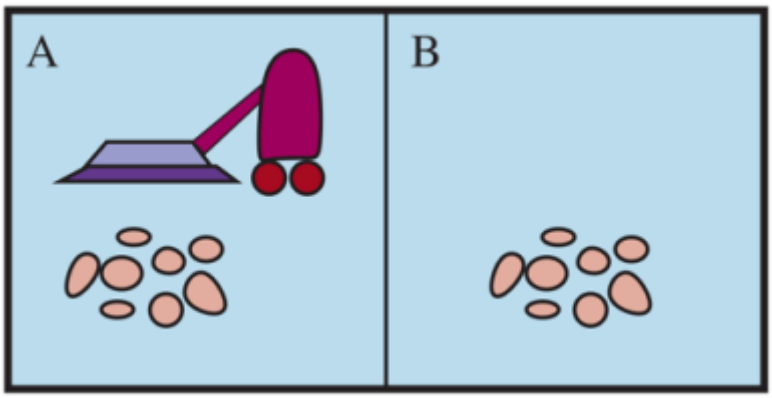
Agente que aprende:

Estos agentes tienen la capacidad de mejorar su eficacia usando mecanismos de aprendizaje compuestos por:

- Elemento de actuación: encargado de seleccionar las acciones a realizar.

- Elemento de aprendizaje: modifica el elemento de actuación
- Crítica: retroalimentación para el elemento de aprendizaje. Da feedback sobre la actuación del agente para así determinar cómo modificar el elemento de actuación.
- Generador de problemas: sugiere acciones que harán que el agente pruebe experiencias nuevas e informativas que puede que no sean las más útiles a corto plazo pero pueden si serlo en el largo plazo.

5. Defina una función que determine la medida de rendimiento para el ambiente de la aspiradora descripto en la figura 2.2 del libro "Inteligencia Artificial" que se muestra abajo:



La medida de rendimiento será una función que busca maximizar la limpieza y minimizar el costo de tiempo y energía.

Se define como:

1. +10 puntos por cada unidad de suciedad aspirada.
2. -2 puntos por la acción 'Aspirar' (para sumar el gasto de electricidad).

El puntaje total acumulado tras un período de tiempo determinado es la medida de rendimiento del agente.

Un agente racional será aquel que, en un periodo de tiempo determinado, intente maximizar este puntaje total.

6. Examine ahora la racionalidad de las siguientes funciones de agentes aspiradora.

a- Muestre que la función de agente aspiradora descrita en la Fig.2.3 es realmente racional bajo la hipótesis presentada en la sección 2.2 Buen Comportamiento.

Figure 2.3

Percept sequence	Action
$[A, Clean]$	*Right*
$[A, Dirty]$	*Suck*
$[B, Clean]$	*Left*
$[B, Dirty]$	*Suck*
$[A, Clean], [A, Clean]$	*Right*
$[A, Clean], [A, Dirty]$	*Suck*
⋮	⋮
$[A, Clean], [A, Clean], [A, Clean]$	*Right*
$[A, Clean], [A, Clean], [A, Dirty]$	*Suck*
⋮	⋮

La sección 2.2 define a un agente racional de la siguiente manera:

<<Un agente racional es aquel que hace las cosas bien. Obviamente, hacer las cosas bien es mejor que hacer las cosas mal, pero,¿ A qué se refiere con hacer las cosas bien?.>>

Con esta definición en mente podemos decir que la función es realmente racional, ya que vemos que si la casilla en la que está parado está sucia, el agente aspira inmediatamente, lo cual aumenta el rendimiento porque limpia una celda en el menor tiempo posible. En cambio, si la casilla está limpia, la acción que realiza para maximizar el rendimiento es moverse a la otra casilla para verificar si está sucia, ya sea a la derecha o a la izquierda según su posición.

Por lo tanto, el agente está “haciendo el bien” ya que genera una secuencia de pasos que consigue el puntaje más alto según la medida de rendimiento definida. Como el entorno es totalmente observable, determinista y estático, las reglas del agente son suficientes para garantizar la racionalidad.

b- Describa una función para un agente racional cuya medida de rendimiento

modificada deduzca un punto para cada movimiento. ¿Requiere el programa de agente estado interno?

La función sería la siguiente:

+10 Puntos por la acción Limpiar

-1 Punto por movimiento (Derecha o Izquierda)

En un entorno conocido totalmente estático (sin la posibilidad de que se vuelvan a ensuciar casillas ya aspiradas) nuestra aspiradora luego de limpiar ambas casillas se quedaría oscilando entre el movimiento Derecha e Izquierda. Por ende, sería un agente tonto y muy poco racional, ya que la nueva función definida le resta puntos con cada movimiento y el agente reactivo no tiene forma de darse cuenta de que ya visitó ambas casillas. En cambio, si tuviéramos un estado interno que guarde las casillas que se encuentren limpias, tendríamos un agente mejorado que, en vez de perder puntos sin hacer nada, maximizaría la ganancia ya que en caso de no encontrar espacios conocidos sucios dicha aspiradora se quedaría en la acción definida como NoOp (No operar) haciendo que no exista ninguna pérdida de puntos por movimientos innecesarios.

Una función que efectúa lo comentado sería:

```
función AGENTE-ASPIRADORA-CON-ESTADO([localización, estado], memoria) devuelve una acción

        si estado = Sucio entonces

                memoria[localización] ← Limpio

                devolver Aspirar

        En caso contrario, si existe alguna celda C tal que memoria[C] = Desconocida entonces

                si localización = A entonces devolver Derecha

                si localización = B entonces devolver Izquierda

        En caso contrario, si memoria[A] = Limpio y memoria[B] = Limpio entonces

                devolver NoOp
```

c- Discuta posibles diseños de agentes para los casos en los que las cuadrículas limpias puedan ensuciarse y la geografía de medio sea desconocida. ¿Tiene sentido que el agente aprenda de su experiencia en estos casos? ¿Si es así, que debe aprender?

En este caso, el entorno ya no es estático, porque puede darse que una casilla que el robot ya había limpiado vuelva a ensuciarse. Así, ya no nos alcanza con un agente que simplemente reaccione a la suciedad.

Dos opciones que pueden servirnos para resolver un problema así serían:
Agente basado en modelo: Podría mantener un estado interno que represente lo que no observa

directamente. Si guardara un mapa estimado del entorno recordando qué celdas visitó y su estado previo (limpio/sucio) y actualizara ese modelo cada vez que recibe nueva información, podría planificar movimientos, decidir cuándo revisitar celdas y razonar sobre partes del entorno que no está observando en ese momento.

Agente con aprendizaje: Aún mejor que el planteado previamente, este modelo tendría un estado interno y la capacidad de aprender patrones del entorno para mejorar su política de limpieza. Podríamos tener, por ejemplo, una grilla con el mapa descubierto y con el estado estimado de la suciedad por celda y luego, el modelo haría estimaciones sobre cómo reaparece la suciedad, intentando conseguir estimaciones por celda y por tiempo. Con esto, generaría secuencias para cubrir el mapa, asegurándose de revisitar más seguido aquellas celdas que tienen más probabilidad de estar sucias y optimizando sus rutas.

7. Identifique la descripción REAS que define el entorno de trabajo para cada uno de los siguientes agentes:

a- Robot que juega al fútbol

b- Agente para comprar libros de internet

a) Robot que juega al fútbol

Rendimiento:

El objetivo principal del agente es ganar el partido, lo que implica maximizar la cantidad de goles convertidos y minimizar los recibidos. Además, debe mantener un comportamiento eficiente, evitando penalizaciones y utilizando la energía disponible de manera óptima. Otras medidas de rendimiento incluyen la precisión en los pases, la coordinación con los compañeros, la estabilidad durante el movimiento, y la capacidad de reacción ante eventos imprevistos del juego. También puede considerarse el tiempo de respuesta y la durabilidad del hardware como indicadores secundarios de rendimiento.

Entorno:

El entorno está conformado por la cancha, los límites físicos, la pelota, los compañeros, los rivales, el árbitro y las condiciones externas (como la fricción del suelo o la iluminación). Es un entorno dinámico, parcialmente observable y altamente incierto, ya que los demás agentes también actúan de manera autónoma y simultánea. La interacción entre agentes genera un entorno competitivo y cooperativo a la vez, donde las estrategias y tácticas cambian

constantemente. Además, el entorno incluye factores físicos como colisiones, pérdidas de visión o errores de sensores.

Actuadores:

El robot cuenta con actuadores que le permiten desplazarse, girar, acelerar o frenar mediante motores eléctricos. También dispone de un mecanismo de pateo o empuje de la pelota, y en algunos casos sistemas de comunicación inalámbrica para coordinar estrategias con sus compañeros. Estos actuadores deben ser precisos y responder en tiempo real a las órdenes generadas por el sistema de control del agente.

Sensores:

Sus sensores incluyen cámaras RGB o de profundidad, sensores de distancia (ultrasonido o infrarrojo), acelerómetros, giroscopios e incluso sensores de contacto. A través de ellos, el agente obtiene información sobre la posición de la pelota, la ubicación de otros jugadores y su propio estado (velocidad, orientación, equilibrio). En algunos casos se complementa con GPS o balizas para localización absoluta dentro del campo de juego.

b) Agente para comprar libros por Internet

Rendimiento:

El agente debe cumplir con el objetivo de adquirir el libro solicitado por el usuario de la forma más eficiente posible. Esto implica encontrar el libro correcto al menor costo total (precio + envío), en el menor tiempo de entrega y asegurando que la transacción se realice sin errores. Otras métricas de rendimiento pueden ser la satisfacción del usuario, la confiabilidad de los sitios elegidos y la capacidad del agente de aprender de compras previas para mejorar futuras recomendaciones. El rendimiento también se puede evaluar según la cantidad de operaciones exitosas y la precisión al interpretar las preferencias y pedidos del usuario.

Entorno:

El entorno del agente está formado por los diferentes sitios de comercio electrónico, APIs de tiendas, bases de datos de libros y sistemas de envío. Es un entorno dinámico, ya que los precios y la disponibilidad de stock pueden cambiar en cualquier momento. Además, es parcialmente observable: el agente no siempre tiene acceso completo o actualizado a toda la información, por ejemplo, puede desconocer si un artículo se agotará mientras realiza la compra. También es un entorno multiagente, donde interactúan usuarios, tiendas y servicios de pago, teniendo cada una sus propias reglas de negocio y políticas.

Actuadores:

El agente actúa mediante la ejecución de solicitudes HTTP o llamadas a APIs para realizar búsquedas, filtrar resultados, agregar productos al carrito, iniciar sesiones, completar pagos y enviar notificaciones al usuario. También puede interactuar con formularios web o interfaces gráficas automatizadas según el diseño del entorno. Sus actuadores, en este caso, son las acciones de comunicación digital que modifican el estado del entorno virtual.

Sensores:

Los sensores del agente son los mecanismos mediante los cuales recibe información del entorno. Esto incluye la lectura de respuestas web (HTML o JSON), los resultados de búsqueda de los sitios, las confirmaciones de compra, el estado de los envíos y la detección de errores o interrupciones en el proceso. También toma información del perfil del usuario (dirección, preferencias, métodos de pago, historial) para ajustar su comportamiento y lograr una experiencia más personalizada.

8. Para cada uno de los tipos de agente enumerados en el ejercicio anterior, caracterice el Entorno de acuerdo con las propiedades dadas en la sección 2.3 del libro determinístico vs. Estocástico, Observable vs. No-observable, etc.) y seleccione un diseño de agente adecuado.

Propiedad	Robot que juega al fútbol	Agente comprador de libros
Observabilidad	Parcialmente observable: no puede ver todo el campo simultáneamente y existen oclusiones y ruido en los sensores.	Parcialmente observable: la información de stock, precios o tiempos de envío puede no estar actualizada.
Determinismo	Estocástico: las acciones no garantizan siempre el mismo resultado debido a las condiciones del entorno y a la interacción con otros agentes.	determinista, aunque pueden existir eventos aleatorios como fallos de red o cambios inesperados en el precio o disponibilidad.

Episodicidad	Secuencial: cada acción (pase, movimiento, tiro) influye directamente en el estado futuro del juego.	Secuencial: las decisiones se encadenan en un flujo de pasos hasta concretar la compra.
Dinamismo	Dinámico: el entorno cambia constantemente por las acciones de los rivales, el movimiento de la pelota y el tiempo.	Dinámico: el entorno virtual cambia con frecuencia, afectando precios y disponibilidad.
Discreción	Continuo: tanto el espacio físico como las variables de movimiento son continuas.	Discreto: las acciones se limitan a opciones definidas (buscar, filtrar, etc).
Número de agentes	Multiagente: intervienen varios robots, tanto aliados como oponentes.	Multiagente: interactúa con servidores, tiendas, usuarios y servicios externos.
Conocimiento del modelo	Parcial: conoce las reglas del juego y su propio modelo físico, pero no controla todas las variables del entorno.	Parcial: conoce las APIs y reglas de interacción, pero no tiene control total sobre el mercado o los cambios en las tiendas.

Diseño de agente recomendado

Robot que juega al fútbol

El diseño más apropiado es el de un agente híbrido, que combina componentes reactivos, basados en modelo, objetivos y aprendizaje.

La parte reactiva permite responder con rapidez ante eventos inmediatos (evitar choques, interceptar la pelota).

El modelo interno mantiene una representación del campo, la posición de los jugadores y el estado del partido.

El componente basado en objetivos decide acciones que aumentan la probabilidad de alcanzar la meta (marcar un gol o defender).

Finalmente, el aprendizaje permite mejorar el rendimiento a partir de la experiencia, ajustando estrategias y tácticas de equipo

Agente comprador de libros

El diseño más adecuado es el de un agente basado en modelo y utilidad, capaz de razonar sobre los posibles cursos de acción.

El modelo interno del agente incluye la información de los sitios web, precios, costos de envío y políticas de devolución.

El componente de utilidad evalúa y compara diferentes opciones según criterios del usuario (precio, tiempo, reputación).

Puede incluir un módulo de aprendizaje que personalice el comportamiento según las compras anteriores.

El agente también debe ser capaz de planificar ante cambios en el entorno, como la falta de stock o errores en el pago.