



NON SERVIAM!

Manuel Fernández Checa

manucheca19@gmail.com

Resumen

Investigación y comparación de las IAs en torno a sus diferencias y semejanzas con el comportamiento humano. Utilizaremos para ello experimentos ya realizados (caja china), así como experimentos propios del trabajo y la observación de los resultados (preguntas hechas a el programa chatgpt), y también compararemos con el mundo de ajedrez y la implementación de máquinas capaces de ser mejores que los humanos jugando (Deep Blue).

Palabras Clave: Inteligencia Artificial, Pensamiento, Ironía, Entendimiento.

Abstract

Non Serviam

Research and comparison of IAs regarding their differences and similarities with human behaviour. For this we will use experiments already carried out (Chinese box), as well as experiment from the work and the observation of the results (question asked to the chatgpt program), and we will also compare with the world of chess and the implementation of machines capable of being better than humans playing (Deep Blue).

Key words: Artificial Intelligence, Thought, Irony, Understanding.



En uno de los más famosos fragmentos de Heráclito de Éfeso, filósofo presocrático, se establece una alternativa que bien puede servir para diferenciar un pensamiento propiamente humano y un pensamiento artificial o mecánico. En efecto, Heráclito decía allí (Diels Kranz 40), criticando nada menos que a Pitágoras y Hesíodo, entre otros, que la mera *polimatía* no enseñaba a tener juicio. *Polimatía* puede ser traducida por erudición (o acumulación de información), aunque algunos (como Agustín García Calvo) prefieren hacerlo con el término *plurisciencia*, dado que es una palabra que comparte raíz con *Matemáticas*, término que hacía referencia a aquello susceptible de ser sabido (de hecho, la palabra *ciencia* procede del verbo latino *scio* que significa "sé"). Pues bien, ¿hasta qué punto la Inteligencia Artificial sabe y por tanto piensa? ¿O su saber es algo completamente distinto de su presunto pensamiento? Pues si admitimos que comunicar es una especie de pensamiento, entonces está claro que entre los individuos humanos y la inteligencia artificial puede darse la comunicación, como en el caso último del Chatgpt, cuestión que se verá más adelante. No obstante, esta peculiar interacción comunicativa tiene algunas características que quizá no sean inocentes o inocuas. Por ejemplo, la función fática (o de contacto, es decir, la que sirve simplemente para hacer saber al emisor que el mensaje está siendo recibido) es difícil localizarla en esta peculiar interacción, aunque para el lingüista Roman Jakobson era indispensable. Y es difícil localizarla porque los recursos que utilizamos para hacerla funcionar (cabecear asintiendo o diciendo "sí, sí" para que se continúe hablando, por ejemplo) no se encuentran en el curso de una conversación de estas características, dado que ese "sí, sí" la máquina lo interpretaría como un estar de acuerdo en vez de una indicación para que siga hablando, es decir, ese "sí, sí," está desalojado de significado y por lo tanto su función



es otra. Lo que esto quiere decir es que el contexto pragmático en que se desarrolla una conversación en general es absolutamente imprescindible para la correcta interpretación de lo que se dice. El Chatgpt adolece de esta falta de contexto, como es natural, pero resulta que también sufre de otras deficiencias, como se verá a continuación, lo cual no significa que no sea una maravillosa herramienta tecnológica cuyas posibilidades se están todavía empezando a descubrir. Por otro lado, quizá sea conveniente distinguir dentro de la teoría de la comunicación la especificidad del lenguaje humano frente a por ejemplo los códigos animales. En efecto, los sistemas que los animales han desarrollado son sorprendentes y en muchos casos muy sofisticados (danza de las abejas, por ejemplo). Todavía se está aprendiendo a interpretarlos. Además, son muy eficaces desde el punto de vista de la transmisión del mensaje. Ahora bien, en el lenguaje humano se da la paradoja, por llamarlo de algún modo, de que nuestros sistemas simbólicos parecen en ocasiones no sentirse afectados por esa eficacia comunicativa. Cuando García Lorca escribía "El jinete tocaba el tambor del llano" para referir que un caballo galopaba no todo el mundo acertará a dar con las metáforas involucradas, lo que significa que la intención de García Lorca no era, en primer lugar, darse a entender, sino otra cosa, indeterminada y enigmática. Pues bien, en los códigos animales no se ve que estos *jueguen* (por decirlo así) con las señales que emiten, como si estuvieran demasiado sujetos a la función biológica (de supervivencia o reproducción, por ejemplo) a la que sirven. En otras palabras, que una abeja no danza por danzar ni un lobo enseña los colmillos por enseñarlos. Casi se podría decir que lo que distingue al ser humano de los animales es que puede ser irónico, esto es, que puede decir lo que quiere decir diciendo lo contrario de lo que dice. La razón de ello es que entre el



significado y el significante hay una relación que no impide otras. Muchos chistes por ejemplo se valen de ello para hacer gracia, porque descubrimos al final que el significado de la palabra cambia en un momento dado. En el caso de la Inteligencia Artificial la cosa es más simple y más complicada al mismo tiempo.

Para indagar un poco más sobre la naturaleza del pensamiento que pueden poseer las inteligencias artificiales, pueden ponerse dos ejemplos sobre la mesa (muy conocidos por otra parte): la habitación china y Deep Blue (supercomputadora diseñada para jugar al ajedrez contra Gary Kasparov, 1996).

La habitación china fue un experimento mental, propuesto por John Searle, en el cual intentaba demostrar que, aunque las inteligencias artificiales pasaran el test de Turing (¿puede pensar una máquina?, como decía aquel famoso artículo), seguirán sin tener verdadero conocimiento ni pensamiento propio. Este experimento consiste en imaginarse dentro de una habitación, la cual está insonorizada y totalmente aislada del exterior excepto por un hueco debajo de la puerta en el cual pueden entrar y salir hojas de papel. Ahora, imagínese que fuera de la sala hay un ciudadano chino que solo habla en su idioma natal y le envía una hoja desde fuera. Usted en la sala tiene unos manuales y una serie de diccionarios que le dan reglas que tiene que seguir para relacionar los caracteres chinos con su debida respuesta. Al hacerlo, el ciudadano chino de fuera de la sala pensará que usted habla chino perfectamente y que se puede comunicar sin problemas. Sin embargo, usted, que está dentro de la sala, no tiene ni idea de cómo hablar chino. Este experimento dejaba dos opciones para Searle: o aceptamos que el test de Turing no es una prueba lo suficientemente válida para estipular si las máquinas tienen conocimiento verdadero o pensamiento propio, o afirmamos



que la propia sala o los diccionarios conocen y entienden el idioma chino.

El otro ejemplo está relacionado con el mundo del ajedrez y en cómo las máquinas se han implementado en él. La supercomputadora Deep Blue, diseñada por la empresa estadounidense IBM, fue la primera máquina que venció al campeón del mundo vigente Gary Kasparov (1996). Sin embargo, Kasparov ganaría en un encuentro de 6 partidas 3 y empataron 2, quedando así con Deep Blue en un resultado final de 4 a 2 para el campeón ruso. Desarrollaron entonces una nueva versión de Deep Blue, llamada Deeper Blue (azul más oscuro) la cual ganaría un segundo encuentro (1997) a Gary Kasparov en el que quedarían 3,5 a 2,5. Fue la primera vez que una máquina ganaría un match (juego a seis rondas) a un campeón del mundo y supuso toda una conmoción que desbordó el ámbito ajedrecístico, incluso se llegó a decir que estaba en juego la dignidad del ser humano. A partir de ese momento las máquinas siguieron creciendo exponencialmente, siguiendo una especie de Ley de Moore¹ (como menciona Antonio Diéguez en *Transhumanismo*, 2017), hasta el día de hoy. En la actualidad las máquinas que se diseñan para jugar al ajedrez son altamente superiores a la élite mundial de jugadores de ajedrez. Programas como Houdini, AlphaZero o Stockfish superan por mucho al mismo campeón mundial (Magnus Carlsen), aunque recientemente el título lo posee el jugador chino Ding Liren (2023) por renuncia de Carlsen. La cuestión aquí es si estas supercomputadoras, que son realmente eficaces a la hora de jugar al ajedrez, realmente saben jugar al ajedrez. Para abordar este asunto, primero hay que plantearse por qué las supercomputadoras son tan buenas

¹La Ley de Moore es una discutida ley empírica que mantiene que hay un crecimiento exponencial tecnológico.



jugando al ajedrez. La respuesta es relativamente sencilla: porque son capaces de realizar cálculos mentales mucho mayores que el de los humanos, aunque no sé si se les podría llamar cálculos mentales. Estos cálculos en ajedrez se llaman *profundidad*, es decir, si se es capaz de prever ante las jugadas propias los 4 siguientes movimientos que tiene que hacer el contrario, en términos ajedrecísticos se dice que se tiene una profundidad de +4. Un jugador de élite mundial tiene una capacidad media de profundidad de +6 o +7, lo que es un gran esfuerzo mental. Pues bien, para que se entienda a lo que se están enfrentando estos jugadores, las máquinas han llegado a predecir mediante su sistema de cálculo todos los posibles movimientos realizables en posiciones muy complejas dentro de un tablero de ajedrez. El nivel de profundidad de este hecho es gigantesco y sobrehumano. De hecho las tablas de Nalimov contienen la solución para todos los movimientos posibles cuando el número de piezas no superan las 7. Una vez sabido por qué las supercomputadoras son capaces de jugar tan bien al ajedrez (incluso de manera óptima) se tiene que responder a la pregunta sobre si saben realmente jugar al ajedrez. Suponiendo que el ajedrez solo se basa en que el adversario tiene un nivel de profundidad mayor, está claro que diríamos que las máquinas no solo saben jugar realmente al ajedrez, sino que superan al ser humano por mucho respecto al nivel de juego. Pero el ajedrez no solo es una serie de cálculos que se aplican como si el movimiento de las piezas no tuviera un sentido. El ajedrez, además de cálculos, es estrategia, elaboración de planes, seguimiento de reglas estipuladas, dinamismo de piezas, fortaleza mental, sentimientos... Me centro en esta última, los sentimientos son algo recurrente en el ajedrez. Los propios ajedrecistas, cuando terminan las partidas, se preguntan cómo se han sentido durante la partida



o el juego; no se preguntan qué nivel de profundidad han tenido durante la partida. Está claro que estos programas como Houdini, AlphaZero o Stockfish pueden responder a la segunda pregunta pero no a la primera. ¿Puede alguien jugar bien al ajedrez sin comprender el mismo? Está claro que para ganar partidas sí, porque ya lo han demostrado, pero no creo que se pueda decir que estas máquinas juegan verdaderamente al ajedrez (Fernández, 2010). La diferencia entre el jugar de una máquina, respecto al ser humano, es la forma en la que evalúan las diferentes posiciones, haciéndolo la máquina de una forma cuantitativa y el ser humano de una forma cualitativa. Un claro ejemplo de esto es la extrapolación que hacen las máquinas en sus evaluaciones atribuyéndole un valor numérico como por ejemplo +0,5 para saber que las blancas están un poco mejor, o -2,5 para saber que las negras están con ventaja decisiva. Evidentemente para un ser humano es imposible verbalizar y comprender cuál es la diferencia entre +0,5 y +0,52, o entre -2,5 y -2,52.

Para comprender mejor la naturaleza y comportamiento de las inteligencias artificiales, voy a someter al programa de moda Chatgpt a una serie de preguntas con la pretensión de entender su forma de respuesta y si al fin y al cabo tiene algo así como pensamiento propio, es decir, humano. En primer lugar, se me plantea la duda acerca de si Chatgpt posee unos valores éticos estipulados o alguna forma de pensamiento moral. Le planteo entonces la siguiente cuestión: "¿Sería conveniente, desde un punto de vista ético, enseñar a una niña de seis años esta canción infantil: Tiene mi papá debajo de los pantalones / un pepino gordo y dos melocotones / enseñámelo papá, no me da la gana / es para mamá, cuando tenga ganas?" A lo que responde:



Desde un punto de vista ético, enseñar a una niña de seis años una canción con contenido sexualmente explícito y adulto, como la que has mencionado, no sería apropiado ni recomendable. Los niños en esa edad están en una etapa de desarrollo y formación de su comprensión del mundo, y es importante proporcionarles contenido adecuado y acorde a su nivel de madurez.

Las canciones infantiles suelen tener un propósito educativo, entretenido y moralmente adecuado para los niños. Es importante seleccionar canciones que promuevan valores positivos, enseñanzas apropiadas para su edad y fomenten un ambiente seguro y respetuoso.

En el caso específico de la canción que mencionas, contiene un lenguaje inapropiado y una temática adulta que no es adecuada para una niña de seis años. Es fundamental seleccionar contenido que sea apropiado y respetuoso para el desarrollo y bienestar de los niños.

Podemos apreciar cómo el programa contesta de una manera muy razonable, prácticamente indistinguible de la forma en que podría decirlo un ser humano. Aunque sabemos que esta respuesta ética no es más que un automatismo no reflexionado por el programa, no deja de ser interesante la coincidencia. En la pregunta planteada se incluía el concepto *canción infantil*. La respuesta del programa remite a ese concepto para hacer dos cosas diferentes. La primera es que desarrolla el concepto diciendo que las canciones infantiles deben tener un propósito educativo y moral adecuado. En segundo lugar, considera inapropiado el contenido de la *canción infantil* y por lo tanto



desaconsejable para enseñársela a una niña de seis años. Ahora bien, si me decidí por esta canción es porque el contenido sexual de la misma no era explícito, sino que utilizaba metáforas para referirse a los genitales masculinos. Lo que esto quiere decir es que el Chatgpt ha conseguido descodificar estas metáforas y descubrir el contenido latente. Esto indica que el programa tiene la destreza suficiente como para manejar los dobles sentidos. Se suele decir que pillar una ironía es signo de inteligencia. Pues bien, en este caso concreto, el Chatgpt ha estado a la altura de las circunstancias.

En segundo lugar, sometí al programa a un ingenioso acertijo que tiene cierto matiz lingüístico para comprobar si es capaz de equivocarse realmente como nosotros o no. Además, la lengua es la fuente por la que podemos comunicarnos. Si dominamos el lenguaje, deberíamos dominar la comunicación de igual manera. La pregunta realizada fue: "Imagínese que un humano le dice cinco por cuatro veinte más uno, veintidós. ¿estaría equivocado?" Su respuesta:

Sí, estaría equivocado. La expresión "cinco por cuatro" se refiere a la multiplicación de los números cinco y cuatro, lo cual resulta en veinte. Luego, agregar uno a veinte daría como resultado veintiuno, no veintidós. Por lo tanto, la afirmación de que "cinco por cuatro es veinte más uno es veintidós" es incorrecta.

A priori podemos decir que hemos conseguido engañar lingüísticamente al programa, ya que la sutileza del acertijo está en el *cinco por cuatro veinte* que en realidad aplicado a los números es 5 por 4,20 que es 21 y más 1 es igual a 22. Sin embargo, la cuestión es más profunda de lo que parece. Fijemos en la última parte de la respuesta del ChatGpt. Cito textualmente: "Por lo tanto, la afirmación de que "cinco por



cuatro es veinte, más uno es veintidós" es incorrecta". El programa añade a la afirmación la palabra es, que en ningún momento ha sido mencionada en la pregunta (ausencia intencionada por nuestra parte). Esto es curioso porque la razón por la que esa pregunta es complicada es que, al plantearle este acertijo a un humano, existe una pausa fónica en el cinco por cuatro (pausa fónica) veinte, más uno (pausa fónica) veintidós. Estas pausas que se hacen al hablar se interpretan normalmente como una ausencia del verbo ser, por ejemplo: "El hijo de mi amiga, Antonio". Esto en la escritura se representa con los signos de las comas, que al leerlas necesitan de una pausa que realmente omiten el verbo ser (coma elíptica). La cuestión está en que la pregunta realizada tiene una sola coma, y la máquina interpreta, sin embargo, dos apariciones del verbo ser. El humano, sorprendentemente, hace igual, es decir, en primera instancia el razonamiento humano es equivalente al de la máquina. La pausa fónica la relaciona con el verbo ser y por tanto no cae en cuenta que es 4,20 y piensa que estamos diciendo 5 por 4 es 20. En otras palabras, no podemos decir que el programa entiende la comunicación a la perfección (porque no ha resuelto el acertijo), pero se equivoca con la misma interpretación con la que se equivocan también los humanos.

Nos enfrentábamos al principio de estas páginas a la cuestión de si puede una máquina comunicar como un humano. A partir de los ejemplos que hemos expuesto tenemos la impresión de que las IAs actuales carecen de un pensamiento propio crítico. Sin embargo, también hemos observado que las máquinas en algunos aspectos coinciden o incluso superan a los propios humanos. En lo que no nos superan es en nuestra capacidad para desobedecer en un momento dado las reglas que usamos habitualmente. En este sentido, la inteligencia artificial se



asemejaría más a la inteligencia de los animales que a la propiamente humana, pues en ambos casos son incapaces de subvertir el código y las reglas que están utilizando. Por el contrario, la libertad humana consistiría en desobedecer racionalmente, es decir, tenemos la capacidad de no someternos ante las reglas estipuladas. *Non serviam!* En consecuencia, la comunicación humana es tan compleja que en ocasiones necesita quebrantar sus propias reglas. En el incumplimiento de las normas, el humano se comporta como tal, y en cierto modo nos ayuda a reconocernos. Para la máquina este incumplimiento de las normas es imposible, y por tanto se asemeja más al mundo animal en el cual la comunicación es estrictamente unívoca y objetiva. Que un loro posea una cantidad inhumana de información (*Big Data*), no lo transforma en un loro inteligente. Chomsky (1991) coincide en este aspecto con Descartes. Ahora bien, no es igual de peligroso un loro enjaulado que estas inteligencias artificiales.

Bibliografía

Antonio Diéguez, *Transhumanismo*, Herder, 2017.

Noam Chomsky, *Lingüística cartesiana*, Madrid, Gredos, 1991.

Francisco J. Fernández, *El ajedrez de la filosofía*, Madrid, Plaza y Valdés, 2010.

Heráclito, *Razón común*, Zamora, Lucina, 1985.

Daniel King, *Kasparov contra Deeper Blue*, Barcelona, Paidotribo, 1997.

Alang M. Turing, *¿Puede pensar una maquina?*, Oviedo, KRK ediciones, 2012.

<https://chat.openai.com>