¿PUEDEN PENSAR LAS MÁQUINAS?

A.M. Turing, *Computing Machinery and Intelligence*

EL JUEGO DE LA IMITACIÓN

Me propongo examinar la pregunta siguiente: "¿Pueden pensar las máquinas?". Deberíamos empezar con definiciones acerca del significado de los términos "máquina" y "pensar". Las definiciones podrían ser elaboradas para reflejar, lo mejor posible, el uso normal de las palabras, pero tal posición es peligrosa. Si quisiéramos encontrar el significado de las palabras "máquina" y "pensar" examinando cómo suelen ser empleadas dichas palabras, sería difícil escapar a la conclusión de que habría que buscar el significado y la respuesta a la pregunta: "¿Pueden pensar las máquinas?", en una indagación estadística, a la manera de las encuestas de Gallup. Pero esto es absurdo. En vez de intentar tal definición, sustituiré la pregunta por otra que está estrechamente vinculada con ella y es expresada en palabras relativamente claras.

La nueva forma del problema puede ser definida en función de un juego que llamamos "juego de la imitación". Se juega con tres personas, un hombre (*A*), una mujer (*B*) y un interrogador (*C*), que puede ser de uno u otro sexo. El interrogador se queda en un cuarto, separado de los otros dos. Para él, el objetivo del juego es determinar cuál de los otros dos es el hombre y cuál la mujer. Los conoce por los signos *X* e *Y*, y al final del juego puede decir: " *X* es *A* e *Y* es *B*", o bien: " *X* es *B* e *Y* es *A*". El interrogador puede hacer preguntas a *A* y a *B* de la manera siguiente:

C: Que diga *X* cuán largo tiene el pelo.

Supongamos que *X* es en realidad *A*. En este caso, *A* tiene que responder. El fin de *A* en el juego es tratar de lograr que *C* haga una identificación incorrecta. Por lo tanto, su respuesta podría ser la siguiente: "Me cortaron el pelo y mis cabellos más largos tienen aproximadamente veinte centímetros".

Para que el tono de la voz no ayude al interrogador, las respuestas deberían ser escritas. Lo ideal sería que un teleimpresor estuviera en comunicación con ambos cuartos. La pregunta y las respuestas pueden ser alternativamente repetidas por un intermediario. El objeto del juego para el tercer jugador (*B*) es ayudar al interrogador. La mejor estrategia para ella, probablemente será contestar la verdad. Puede añadir a sus respuestas cosas como: "Yo soy la mujer, ¡no le crea a él!".

Eso, sin embargo, no adelantará nada, porque el hombre puede formular observaciones semejantes.

Ahora preguntamos: "Qué pasaría si una máquina tomara el papel de *A* en este juego?". ¿Decidirá el interrogador equivocadamente con igual frecuencia, si el juego se desarrolla de esta manera, que como cuando es jugado por un hombre y una mujer? Estas preguntas sustituyen la nuestra anterior: "¿Pueden pensar las máquinas?".

CRÍTICA DEL NUEVO PROBLEMA

Al igual que preguntar: "¿Cuál es la respuesta a esta nueva forma de la pregunta?", podríamos preguntar: "¿Vale la pena investigar esta nueva pregunta?". Esta última pregunta la examinamos sin más rodeos, cortando así un regreso sin fin.

El nuevo problema tiene la ventaja de trazar una línea bastante bien definida entre las capacidades físicas e intelectuales del hombre. Ningún ingeniero o químico alega que es capaz de producir un material indistinguible de la piel humana. Es posible que un día sea posible hacer esto, pero incluso si tal invención fuese asequible, pensaríamos que tendría poco sentido tratar de hacer una "máquina que piensa" más humana, cubriéndole con un tejido artificial de este tipo. La forma en que hemos planteado el problema refleja este hecho en la condición que impide al interrogador ver o tocar a los demás competidores o

escuchar sus voces. Algunas otras ventajas del criterio propuesto podrían ser demostradas con los siguientes ejemplos y contestaciones:

P: Escríbame un soneto sobre el tema del Puente de Forth, por favor.
R: No cuente conmigo en este caso. Nunca he sabido escribir poesía.

P: Sume 34957 a 70764.
R: (Pausa de aproximadamente treinta segundos. Después, la contestación) 105721.

P: ¿Juega usted al ajedrez?
R: Sí.
P: Tengo R en mi R1 y ninguna otra pieza. Usted tiene solamente R en R6 y T en T1. Le toca jugar. ¿Cómo jugará?
R: (Después de una pausa de quince segundos). T-T8 mate.

El método de la pregunta y contestación parece ser conveniente para introducir casi cualquiera de las esferas de la actividad humana que queremos incluir. No queremos castigar a la máquina por su incapacidad de brillar en competencias de belleza, ni castigar a un hombre porque pierde una carrera contra un avión. Las condiciones de nuestro juego ponen esas incapacidades fuera de lugar. Los "testigos", si lo consideran oportuno, pueden jactarse cuanto quieran de sus encantos, fuerza o heroísmo, pero el interrogador no puede pedir demostraciones prácticas.

El juego podría, tal vez, ser criticado en el sentido de que la máquina tiene que luchar contra una fuerza demasiado superior. Si el hombre tratara de fingir que él es la máquina, evidentemente haría un mal papel. Sería descubierto inmediatamente por su lentitud e inexactitud en la aritmética. ¿No es posible que las máquinas realicen algo que debería ser definido como pensamiento, pero que es muy diferente de lo que hace un hombre? Esta objeción es muy fuerte. Sin embargo, por lo menos podemos decir que si, no obstante, es posible producir una máquina que pueda jugar satisfactoriamente el juego de la imitación, no necesitamos preocuparnos de esta objeción.

Se podría recomendar que la mejor estrategia para la máquina en el "juego de la imitación" podría ser, tal vez, otra cosa que la imitación de la conducta de un hombre. Eso es posible. Yo pienso, sin embargo, que es poco probable que haya algún gran efecto de este tipo. De todas maneras, no tenemos la intención de investigar la teoría del juego aquí. Vamos a suponer que la mejor estrategia es tratar de dar las mismas respuestas que serían dadas por un hombre.

.....

PUNTOS DE VISTA ANTAGÓNICOS

No podemos abandonar completamente la forma original del problema, porque habrá diferencias de opinión en cuanto a la conveniencia de la sustitución y tenemos que escuchar, por lo menos, lo que será dicho en este respecto.

Será una simplificación del asunto para el lector, si primero explico mis propias opiniones acerca de este asunto. Consideremos primero la forma más precisa de la pregunta. Creo que aproximadamente dentro de cincuenta años será posible programar computadoras que posean una capacidad de almacenamiento de aproximadamente 109, para hacerlas participar en el juego de la imitación tan bien, que un interrogador corriente no tendrá más de un 70% de oportunidad de hacer la identificación correcta después de cinco minutos de preguntas. La pregunta original: "¿Pueden pensar las máquinas?" me parece demasiado carente de sentido para que merezca una discusión. No obstante, creo que hacia fines de este siglo el uso de las palabras y la opinión culta en general habrán cambiado tanto, que será posible hablar de máquinas que piensan sin ser desmentido. Pienso, además, que ocultar estas opiniones no serviría a ningún propósito útil. La opinión popular de que los trabajadores científicos proceden inexorablemente de un hecho bien establecido a otro, sin estar jamás influidos por suposiciones no comprobadas, es muy errónea. Siempre que

sea bien claro cuáles son los hechos probados y cuáles las suposiciones no comprobadas, no puede resultar dañoso. Las conjeturas son muy importantes porque sugieren líneas de investigación científica útiles.

Ahora procederé al examen de opiniones opuestas a las mías.

1) ***La objeción teológica.*** El pensamiento es una función del alma inmortal del hombre. Dios ha dado un alma inmortal a todos los hombres y mujeres, pero a ningún otro animal o las máquinas. Por lo tanto, ningún animal o máquina puede pensar.

No puedo aceptar ninguna parte de esta objeción, pero trataré de contestar en términos teológicos. El argumento me parecería más convincente si los animales fueran clasificados con las personas; para mí, hay más diferencia entre el típico ser animado y el inanimado, que entre el hombre y los otros animales. El carácter arbitrario del criterio ortodoxo es aún más claro si consideramos cómo le podría parecer a un miembro de alguna otra comunidad religiosa. ¿Cuál es la opinión de los cristianos acerca del criterio musulmán de que las mujeres no tienen alma? Pero dejemos este punto y volvamos al argumento principal. Me parece que el argumento citado anteriormente implica una seria restricción de la omnipotencia del Todopoderoso. Se admite que hay ciertas cosas que no puede realizar, como por ejemplo hacer que uno sea igual a dos, ¿pero no deberíamos creer que tiene la libertad de conferir un alma a un elefante, si lo considera adecuado? Suponemos que ejercería este poder solamente en conexión con un cambio que proporcionaría al elefante un cerebro adecuadamente evolucionado, para que correspondiera a las necesidades de su alma. Podríamos argumentar exactamente de la misma forma en el caso de las máquinas. Puede parecer diferente porque es más difícil de "tragar". Sin embargo, esto en realidad sólo quiere decir que nosotros pensamos que sería menos probable que el Omnipotente considerara las circunstancias adecuadas para conferir un alma. Las circunstancias del caso son discutidas en el resto de este ensayo. Al tratar de construir tales máquinas no deberíamos irreverentemente usurpar Su poder de crear almas en ninguna medida mayor que en la que estamos haciéndolo en la procreación de niños: en ambos casos, somos más bien instrumentos de Su voluntad, porque proporcionamos albergues a las almas que El crea.

.....

2) ***La objeción de la "cabeza en la arena".*** "Las consecuencias del hecho de que las máquinas pensaran serían demasiado horribles. Esperemos y creamos que no puedan hacerlo".

Este argumento rara vez es expresado tan abiertamente como en la forma anterior. Sin embargo, afecta a la mayoría de los que no pensamos en el problema en absoluto. Nos gusta creer que el hombre es superior, de una manera sutil, al resto de la creación, y tanto mejor si puede ser demostrado que es necesariamente superior, porque en ese caso no hay peligro de que pierda su posición dominante. La popularidad del argumento teológico está evidentemente relacionada con esta opinión. Suele ser muy fuerte entre los intelectuales, porque atribuyen más valor al poder de pensar que los demás, y tienden a fundar su creencia en la superioridad del hombre en este poder.

No creo que este argumento sea tan sólido que requiera refutación. Un consuelo sería más adecuado: tal vez debería ser buscado en la transmigración de las almas.

3) ***La objeción matemática.*** Hay muchos resultados de la lógica matemática que pueden ser utilizados para demostrar que existen limitaciones al poder de las máquinas de estado discreto. El más conocido de estos resultados se conoce bajo el nombre de Teorema de Gödel y demuestra que, en cualquier sistema lógico suficientemente poderoso pueden ser formuladas proposiciones que no pueden ser demostradas ni refutadas dentro del sistema, a menos que el sistema mismo sea contradictorio. Existen otros resultados, semejantes en algunos aspectos, como los de Church, Kleene, Rosser y Turing. El que más conviene considerar es el último, porque se refiere directamente a las máquinas, mientras que los demás podrían ser utilizados solamente en un argumento relativamente indirecto: por ejemplo, si quisiéramos utilizar el Teorema de Gödel, necesitaríamos, además, tener algún medio de describir sistemas lógicos en función de máquinas, y máquinas en función de sistemas lógicos. El resultado de que se trata se refiere a un tipo de máquina que es esencialmente una computadora digital con una capacidad infinita. Establece que hay ciertas cosas que tal

máquina no puede hacer. Si se equipa para que dé respuestas a preguntas, como en el juego de la imitación, habrá algunas a las cuales o bien dará una respuesta incorrecta, o no dará ninguna, sea cual fuere su tiempo disponible para contestar. Puede, sin embargo, haber muchas preguntas de este tipo, pero las preguntas que no puedan ser contestadas por una máquina pueden ser contestadas satisfactoriamente por otra. Suponemos, naturalmente, que por el momento se trata de preguntas a las cuales la respuesta "sí" o "no" es apropiada, más bien que de preguntas como: "¿Qué piensa usted de Picasso?". Sabemos que las preguntas en las que las máquinas tienen que fallar son del tipo siguiente: "considere que la máquina tiene las siguientes especificaciones... ¿Contestará esta máquina alguna vez *sí* a alguna pregunta?". Hay que sustituir los puntos suspensivos por la descripción de una máquina, en una forma estandarizada.... Cuando la máquina descrita tiene cierta relación comparativamente simple con la máquina que se encuentra bajo interrogación, se puede demostrar que la respuesta será o bien incorrecta o no habrá ninguna. Este es el resultado matemático: Se argumenta que demuestra una incapacidad de las máquinas a la cual el intelecto humano no está sometido.

La respuesta breve a este argumento es que, a pesar de que se ha establecido que hay limitaciones a las capacidades de cualquier máquina particular, solamente ha sido enunciado, sin ninguna clase de pruebas, que el intelecto humano no sufre de tales limitaciones. Sin embargo, no creo que sea posible pasar por alto tan simplemente este punto de vista. Siempre que se haga a una de estas máquinas la pregunta crítica apropiada, y ella dé una contestación definida, sabemos que esta respuesta tiene que ser incorrecta, lo cual nos da cierto sentido de superioridad. ¿Es ilusorio este sentido? Indudablemente es sincero, pero no creo que se le deba atribuir demasiada importancia. Nosotros mismos, demasiado a menudo, damos respuestas incorrectas a ciertas preguntas, para justificar el hecho de que nos sintamos muy satisfechos de tal prueba de falibilidad de parte de las máquinas. Además, en una ocasión como esta, podemos sentir nuestra superioridad solamente respecto a la máquina concreta que hemos vencido con nuestro triunfo insignificante. No se plantearía la cuestión de triunfar simultáneamente sobre todas las máquinas. En resumen, es posible que haya hombres más listos que cualquier máquina dada, pero entonces puede haber otras máquinas más listas, y así sucesivamente. NOTA 1

.....

4) ***El argumento de la conciencia.*** Este argumento está muy bien expresado en la Oración de Lister del año 1949 del profesor Jefferson, de la cual cito lo siguiente: "Hasta que una máquina no sepa escribir un soneto o componer un concierto con base en los pensamientos y las emociones que siente, y no a consecuencia de la caída venturosa de símbolos, no podremos estar de acuerdo en que la máquina pueda ser igual que un cerebro, es decir, que no solamente sepa escribirlos, sino también que sepa que los ha escrito. Ningún mecanismo podría sentir (y no sólo señalar artificialmente, lo cual es una invención fácil) alegría por sus éxitos, tristeza cuando sus válvulas se fundieran, placer al ser adulado y sentirse desgraciado a consecuencia de sus errores, encantado por el sexo, enfadado o deprimido al no lograr lo que desea".

Este argumento parece ser una negación de la validez de nuestro ensayo. Según la forma más extrema de este punto de vista, la única manera en que una persona podría estar segura de que una máquina piensa, sería siendo la máquina y sintiéndose pensar. Entonces uno podría describir estos sentimientos al mundo, pero, naturalmente, nadie se sentiría obligado a hacerle caso. Además, según este punto de vista, la única manera de saber que un hombre piensa consiste en ser este hombre particular. En realidad, es el punto de vista solipsista. Puede ser que sea el punto de vista más lógico, pero dificulta la comunicación de ideas. *A* está inclinado a pensar: " *A* piensa, pero *B* no", mientras que *B* piensa: " *B* piensa, pero *A* no". En lugar de argumentar continuamente acerca de este punto, es habitual respetar la convención cortés de que todos piensan.

.....

No quiero dar la impresión de que creo que no hay ningún misterio en lo que concierne a la conciencia. Hay, por ejemplo, algo de paradójico relacionado con cualquier intento de localizarla. Sin embargo, no pienso que estos misterios tengan forzosamente que estar resueltos antes que podamos contestar la pregunta a la que incumbe a este ensayo. NOTA 2

5) ***Argumentos desde el punto de vista de diferentes incapacidades.*** Estos argumentos tienen la forma siguiente:

"Admito que usted puede compeler a las máquinas a hacer todas las cosas que acaba de mencionar, pero nunca podrá inducir a una máquina a hacer *X*."

Numerosas *X* son sugeridas en este sentido. Ofreceré una selección:

> Ser bueno, fértil en recursos, guapo, amistoso, tener iniciativa, tener sentido del humor, saber distinguir lo bueno de lo malo, cometer errores, enamorarse, disfrutar las fresas con crema, hacer que alguien se enamore de algo, aprender de la experiencia, emplear las palabras correctamente, ser el tema de sus propios pensamientos, tener tanta variedad de comportamiento como un hombre, hacer algo verdaderamente nuevo....

Por lo general, no se ofrece ningún apoyo a estas declaraciones. Creo que la mayoría de los casos se basan en el principio de la inducción científica. Un hombre ha visto miles de máquinas en su vida. De lo que ve en ellas saca varias conclusiones generales. Son feas, cada una es destinada a un propósito muy limitado, cuando se las requiere para un uso un poco diferente resultan inútiles, la variedad del comportamiento de cualquiera de ellas es muy pequeña, etcétera. Naturalmente, su conclusión es que estas son propiedades indispensables de las máquinas en general. Muchas de estas limitaciones están relacionadas con la muy pequeña capacidad de almacenamiento de la mayoría de las máquinas. Supongo que la idea de la capacidad de almacenamiento es ampliada, de alguna manera, para incluir las máquinas que no son máquinas de estado discreto. La definición exacta no importa, ya que en la discusión actual no se pretende ninguna precisión matemática. Hace algunos años, cuando todavía se oía muy poco acerca de computadoras digitales, era posible expresar bastante incredulidad acerca de ellas, cuando se mencionaban sus propiedades sin describir su construcción. Se supone que esto era debido a una aplicación semejante del principio de la inducción científica. Estas aplicaciones del principio son, naturalmente, en gran medida inconscientes. Cuando un niño que se ha quemado teme al fuego y demuestra que lo teme evitándolo, yo diría que está aplicando la inducción científica. (Podría también, naturalmente, describir su comportamiento de muchas otras maneras). El trabajo y las costumbres del género humano no parecen ser material muy adecuado para que se le aplique la inducción científica. Si queremos obtener resultados dignos de confianza tiene que ser investigada una gran parte del espacio-tiempo. Si no es así podría suceder que decidiéramos (como hace la mayoría de los niños ingleses) que todos hablan inglés y que es una tontería aprender el francés.

Sin embargo, hay que hacer ciertas observaciones especiales acerca de muchas de las incapacidades que han sido mencionadas. La incapacidad de saborear fresas con crema puede haber parecido al lector como algo frívolo. Posiblemente se podría construir una máquina que saboreara ese plato delicioso, pero cualquier intento de hacerlo sería necio. Lo importante acerca de esta incapacidad es que contribuye a algunas de las otras incapacidades, por ejemplo, a la dificultad de que surja el mismo tipo de amistad entre un hombre y una máquina como la que surge entre un hombre blanco y otro hombre blanco o entre un hombre negro y otro hombre negro.

La pretensión de que "las máquinas no pueden cometer errores" parece rara. Uno se siente inclinado a replicar: "¿Acaso son peores por ello?" Sin embargo, adoptemos una actitud de mayor simpatía y tratemos de ver lo que se quiere realmente decir con eso. Pienso que esta crítica puede ser explicada en función del juego de la imitación. Se alega que el interrogador podría distinguir la máquina de la persona simplemente dándoles cierto número de problemas aritméticos. La máquina sería descubierta por su absoluta precisión. La respuesta a esto es simple. La máquina (programada para jugar el juego) no trataría de dar las respuestas *correctas* a los problemas aritméticos. Introduciría deliberadamente errores de una manera calculada para confundir al interrogador. Un error mecánico sería probablemente revelado por una decisión inadecuada respecto a qué clase de error aritmético debería cometer la máquina. Incluso esta interpretación de la crítica no es bastante conveniente. Pero no disponemos de espacio suficiente para ir más adelante en esto. Me parece que esta crítica depende de una confusión entre dos tipos de errores. Podríamos llamarlos "errores de funcionamiento" y "errores de conclusión". Los errores de funcionamiento son debidos a cierta falla mecánica o eléctrica que causa que la máquina se porte de otra manera que la proyectada. En las discusiones

filosóficas tendemos a pasar por alto los errores de este tipo; y por eso discutimos de "máquinas abstractas". Estas máquinas abstractas son ficciones matemáticas más bien que objetos físicos. Según su definición, son incapaces de cometer errores de funcionamiento. En este sentido, podemos verdaderamente decir que "las máquinas nunca pueden cometer errores". NOTA 3 Los errores de conclusión pueden surgir solamente si se atribuye alguna importancia a las señales de salida de la máquina. La máquina podría, por ejemplo, mecanografiar ecuaciones matemáticas o frases en inglés. Cuando escribe una proposición falsa, decimos que la máquina cometió un error de conclusión. No hay ninguna razón para decir que una máquina no puede cometer este tipo de error. Podría, por ejemplo, no hacer otra cosa que mecanografiar repetidas veces "0=1". Para tomar un ejemplo menos perverso, podría haber algún método para sacar conclusiones por medio de la inducción científica. Tenemos que suponer que tal método ocasionalmente conduzca a resultados erróneos.

El argumento de que una máquina no puede ser el tema de sus propios pensamientos, puede naturalmente ser contestado solamente si se puede demostrar que la máquina tiene *algún* pensamiento con *algún* contenido. Sin embargo, "el contenido de las operaciones de una máquina" parece significar algo, por lo menos para las personas que tratan con la máquina. Si, por ejemplo, la máquina intentara encontrar una solución a la ecuación: x2-40x-11 = 0, estaríamos inclinados a describir la ecuación como parte del contenido de la máquina en ese momento. En esta clase de sentido una máquina puede indudablemente tener su propio contenido, que puede ser empleado para ayudar a elaborar sus propios programas o para predecir el efecto de los cambios de su propia estructura. Observando los resultados de su propio comportamiento, puede modificar sus propios programas para lograr cierto objetivo más efectivamente. Estas son posibilidades del futuro cercano más bien que sueños utópicos.

La crítica de que una máquina no puede tener mucha diversidad de comportamiento es solamente una manera de decir que no puede tener mucha capacidad de almacenamiento. Hasta muy recientemente, una capacidad de almacenamiento de incluso mil dígitos era muy rara.

.....

6) ***La objeción de Lady Lovelace***. Nuestras informaciones más detalladas acerca de la máquina analítica de Babbage provienen de un informe elaborado por Lady Lovelace. En él declara lo siguiente: "La máquina analítica no pretende *crear* nada. Puede hacer *cualquier cosa que sepamos ordenarle que haga*" (es ella quien subraya). Esta declaración es citada por Hartree, quien añade: "Esto no implica que no sea posible construir equipos electrónicos que `pensarían por sí mismos' o en los cuales, hablando en términos biológicos, sería posible establecer un reflejo condicionado que sirviera de base para aprender. Si esto es posible en principio o no lo es, resulta un problema estimulante y fascinador, insinuado por algunos de los recientes progresos. Sin embargo, no parece que las máquinas proyectadas o construidas en aquel tiempo tuvieran esta propiedad".

Estoy plenamente de acuerdo con Hartree en esto. Se observará que no afirma que las máquinas respectivas no tuvieran esta propiedad, sino que la prueba que Lady Lovelace tenía a su disposición no la alentaba a creer que la tuvieran. Es muy posible que las máquinas respectivas, en cierto sentido, tuvieran esta propiedad. Supongamos que alguna máquina de estado discreto tuviera esta propiedad. La máquina analítica era una computadora digital universal, de manera que, si su capacidad de almacenamiento y su velocidad hubiesen sido adecuadas, habría podido, por medio de una programación adecuada, ser inducida a imitar a la máquina de que estamos tratando. Es probable que este argumento no se le ocurrió a la condesa ni a Babbage. De todas maneras, no tenían ninguna obligación de alegar todo lo que podía ser alegado.

.....

La opinión de que las máquinas no pueden dar origen a sorpresas es debida, creo, a una falacia que cometen particularmente los filósofos y los matemáticos. Es la suposición de que en el momento en que un hecho es presentado a una mente, todas las consecuencias de este hecho se precipitan dentro de la mente al mismo tiempo. Es una suposición muy útil en muchas circunstancias, pero se olvida muy fácilmente que es falsa....

7) ***Argumento de la continuidad en el sistema nervioso***. El sistema nervioso no es, por supuesto, una máquina de estado discreto. Un pequeño error en la información acerca de la dimensión de un impulso nervioso que tropieza con una neurona, puede representar una gran diferencia para el volumen del impulso saliente. Se puede argumentar que, si es así, no se puede esperar que seamos capaces de imitar el comportamiento del sistema nervioso con un sistema de estado discreto.

Es verdad que una máquina de estado discreto tiene que ser diferente de una máquina continua. Sin embargo, si nos adherimos a las condiciones del juego de la imitación, el interrogador no será capaz de sacar ninguna ventaja de esta diferencia. La situación puede ser simplificada si examinamos otra máquina continua más sencilla. Un analizador diferencial servirá muy bien. (Un analizador diferencial es cierto tipo de máquina, no del tipo de estado discreto, que se utiliza para algunos tipos de cálculo). Algunos de estos analizadores presentan sus respuestas en forma mecanografiada, de manera que son convenientes para participar en el juego. No sería posible para una computadora digital predecir exactamente qué respuestas haría el analizador diferencial a un problema; sin embargo, sería capaz de dar el tipo correcto de contestación. Por ejemplo, si se le pidiera que diese el valor de(en realidad, aproximadamente 3.1416), sería razonable escoger al azar entre los valores 3.12, 3.13, 3.14, 3.15, 3.16 con las probabilidades de 0.05, 0.15, 0.55, 0.19, 0.06 (por ejemplo). En estas circunstancias, sería muy difícil para el interrogador distinguir el analizador diferencial de la computadora digital.

8) ***El argumento de la informalidad del comportamiento.*** No es posible elaborar un conjunto de reglas que describa lo que una persona debería hacer en cualquier serie concebible de circunstancias. Puede existir, por ejemplo, la regla de que uno debe detenerse cuando vea una luz roja de tránsito y seguir adelante cuando vea una luz verde; pero ¿y si por alguna falla aparecieron juntas las dos? Tal vez uno decidiera que lo más seguro sería pararse. Sin embargo, más tarde, de esta decisión podría surgir otra dificultad. Intentar proveer reglas de conducta para cada eventualidad, incluso las que provienen de luces para la circulación, parece imposible. Estoy de acuerdo con todo esto.

A base de ellos se argumenta que no podemos ser máquinas. Trataré de reproducir el argumento, pero temo que difícilmente le haré justicia. Parece ser aproximadamente así: "Si cada hombre tuviera una serie definida de reglas de conducta para dirigir su vida, no sería mejor que una máquina. Pero como no existen tales reglas, los hombres no pueden ser máquinas".... No pienso que el argumento sea jamás expresado exactamente así, pero creo, no obstante, que este es el razonamiento que se emplea. Puede, sin embargo, haber cierta confusión entre "reglas de conducta" y "leyes de comportamiento", que oscurezca el caso. Por "reglas de conducta" quiero decir preceptos tales como: "Deténgase si ve luces rojas", según los cuales uno puede actuar y de los cuales uno puede ser consciente. Por "leyes de comportamiento" quiero decir leyes de la naturaleza aplicadas al cuerpo humano, como por ejemplo: "Si usted lo pellizca, chillará". Si sustituyéramos "leyes de comportamiento que regulan su vida" por "leyes de conducta mediante las cuales regula su vida" en el argumento citado anteriormente,... creemos que no solamente es verdad que ser regulado por leyes de comportamiento implica la necesidad de ser cierto tipo de máquina (a pesar de que no indispensablemente una máquina de estado discreto), sino también, a la inversa, que ser tal máquina implica ser regulado por tales leyes. Sin embargo, no podemos convencernos tan fácilmente de la ausencia de leyes completas de comportamiento como de la de reglas completas de conducta. La única manera que conocemos de encontrar tales leyes es la observación científica, y, por cierto, no hemos sabido de ninguna circunstancia en que pudiéramos decir: "Ya hemos investigado suficientemente, no hay tales leyes".

Podemos demostrar con más fuerza que cualquier aseveración de este tipo sería injustificada. Supongamos que pudiéramos estar seguros de encontrar tales leyes si existieran. En tal caso, si tuviéramos una máquina de estado discreto, seguramente sería posible descubrir, observándola, lo suficiente acerca de ella para predecir su futuro comportamiento, y eso durante un período razonable, mil años, por ejemplo. Sin embargo, éste no parece ser el caso. En la computadora Manchester he establecido un pequeño programa con el empleo de solamente 1000 unidades de almacenamiento, por medio del cual la máquina proporcionaba respuestas con un número de dieciséis cifras, cada dos segundos. Quisiera retar a alguien, a aprender de estas respuestas lo suficiente acerca del programa, para ser capaz de predecir cualquier respuesta a los valores que no habían sido ensayados.

.....

MÁQUINAS QUE APRENDEN

El lector habrá creído que no tengo ningún argumento muy convincente de carácter positivo para apoyar mis puntos de vista. Si lo tuviera, no me habría esforzado tanto por señalar las falacias en las opiniones opuestas. Ahora daré la prueba de que dispongo.

Volvamos por un momento a la objeción de Lady Lovelace, quien declaró que una máquina solamente puede hacer lo que le decimos que haga. Se podría afirmar que un hombre puede "inyectar" una idea a una máquina, la cual reacciona en cierta medida, y después cae en la quietud, como una cuerda de piano golpeada con un martillo. Otra comparación sería una pila atómica de dimensiones menores que la crítica: una idea inyectada correspondería a un neutrón que entrara dentro de la pila desde afuera. Cada uno de esos neutrones causará cierta perturbación que gradualmente se desvanece. Sin embargo, si se aumentan suficientemente las dimensiones de la pila, la perturbación causada por tal neutrón que entra desde afuera probablemente seguirá aumentando hasta que toda la pila quede destruida. ¿Hay un fenómeno correspondiente para las mentes? ¿Y hay otra para las máquinas? Parece que hay uno para la mente humana. La mayoría de las mentes humanas parecen "subcríticas", es decir, corresponden, en esta analogía, a las pilas de dimensiones subcríticas. Una idea sometida a tal mente dará, en la mayoría de los casos, por lo menos una idea como respuesta. Una muy pequeña proporción de mentes son supercríticas. Una idea presentada a tal mente puede generar una "teoría" entera, compuesta de ideas secundarias, terciarias y aún más remotas. Las mentes de los animales parecen definitivamente subcríticas. Adhiriéndonos a esta analogía, preguntamos: "¿Una máquina puede ser inducida a ser supercrítica?".

La analogía de la "piel de una cebolla" también es útil. Cuando consideramos las funciones de la mente o del cerebro, encontramos ciertas operaciones que podemos explicar en términos puramente mecánicos. Esto, decimos, no corresponde a la verdadera mente: es un tipo de piel que tenemos que quitar para encontrar la mente verdadera. Pero entonces, en lo que queda, vemos otra piel que hay que despojar, y así sucesivamente. Si procedemos de esta manera, ¿llegaremos finalmente a una piel que no contiene nada? En este último caso, toda la mente es mecánica. (Sin embargo, no sería una máquina de estado discreto. Ya hemos discutido acerca de esto).

No pretendo que estos dos últimos párrafos sean argumentos convincentes. Deberíamos más bien definirlos como "exposiciones que tienden a producir una creencia".

.....

En el proceso de intentar imitar una mente adulta humana, tendemos a pensar mucho acerca del proceso que la llevó al estado en el cual se encuentra. Podemos notar tres componentes:

a) El estado inicial de la mente, es decir, durante el nacimiento.
b) La educación a que ha sido sometida.
c) Otras experiencias, que no podemos definir como educación, a las que ha sido sometida.

En vez de tratar de elaborar un programa para estimular la mente adulta, ¿por qué no tratar más bien de producir un programa para estimular la mente infantil? Si esta estuviera después sometida a un curso apropiado de educación, obtendríamos un cerebro adulto. Presumiblemente, el cerebro infantil es como un cuaderno de apuntes que uno compra en la papelería. Poco mecanismo y muchas hojas en blanco. (Mecanismo y escritura son casi sinónimos, desde nuestro punto de vista). Nuestra esperanza es la de que hay tan poco mecanismo en un cerebro infantil que algo como esto puede ser fácilmente programado. Podemos suponer, como una primera aproximación, que la cantidad de trabajo de educación será muy semejante a la necesaria para un niño humano.

Hemos dividido así nuestro problema en dos partes: el programa infantil y el proceso de educación. Las dos están íntimamente vinculadas. No podemos esperar encontrar una buena máquina infantil en el primer

intento. Hay que ensayar enseñando a una de esas máquinas y ver cómo aprende. Después podemos probar otra y ver si aprende mejor o peor. Hay una conexión evidente entre este proceso y la evolución, mediante las identificaciones

estructura de la máquina infantil = material hereditario
cambios de la máquina infantil = mutaciones
selección natural = juicio del experimentador.

Podemos esperar, sin embargo, que este proceso será más expedito que la evolución. La supervivencia del más apto es un método lento de medir ventajas. El experimentador debería ser capaz de acelerarlo mediante el ejercicio de la inteligencia. El hecho de que no está limitado a mutaciones al azar es igualmente importante. Si puede descubrir la causa de alguna debilidad, será probablemente capaz de pensar en el tipo de mutación que pueda mejorarla.

No será posible aplicar exactamente el mismo proceso de enseñanza a la máquina que a un niño normal. No tendrá por ejemplo, piernas, de manera que no se le podrá pedir que salga fuera para llenar el cubo para carbón. Es posible que no tenga ojos. Sin embargo, por mejor que fueran superadas estas deficiencias mediante una hábil ingeniería, no sería posible mandar la máquina a la escuela sin que los niños se burlaran excesivamente de ella. Hay que darle educación. No necesitamos preocuparnos de las piernas, ojos, etcétera. El ejemplo de Helen Keller demuestra que la educación puede ser proporcionada siempre que por uno u otro medio se establezca comunicación en ambos sentidos entre el maestro y el alumno.

Normalmente, solemos relacionar los castigos y las recompensas con el proceso de la enseñanza. Algunas máquinas infantiles simples pueden ser construidas o programadas con base en este tipo de principio. La máquina tiene que ser construida de tal manera que los resultados que precedieron brevemente a la ocurrencia de una señal de castigo tengan poca probabilidad de ser repetidos, mientras que una señal de recompensa aumenta la probabilidad de la repetición de los acontecimientos que la provocaron. Estas definiciones no presuponen ningunos sentimientos de parte de la máquina. Realicé algunos experimentos con una de esas máquinas infantiles y logré enseñarle algunas cosas, pero el método de enseñanza fue demasiado heterodoxo para que el experimento fuera considerado realmente como un éxito.

.....

Las opiniones acerca de la complejidad adecuada de la máquina infantil pueden variar. Podríamos fabricarla para que fuera lo más simple posible y, al mismo tiempo, de acuerdo con los principios generales. Alternativamente, podríamos tener un sistema completo de inferencia lógica, construido como parte de su estructura. En este último caso, el almacén estaría ocupado en gran medida por definiciones y proposiciones. Las proposiciones tendrían diferentes tipos de nivel, por ejemplo, hechos bien establecidos, conjeturas, teoremas demostrados matemáticamente, aseveraciones hechas por alguna autoridad, expresiones que tienen la forma lógica de una proposición pero ningún valor de fidedignidad. Algunas proposiciones pueden ser definidas como proposiciones "imperativas". La máquina debería estar construida de tal manera que en el momento en que un imperativo estuviese clasificado como "bien establecido" la acción apropiada se efectuase. Para ilustrarlo, supongamos que el maestro dice a la máquina: "Haz tu tarea ahora". Esto puede ser la causa de que la frase: "El maestro dice: `Haz tu tarea ahora"' sea incluida entre los hechos bien establecidos. Otro hecho parecido podría ser: "Todo lo que dice el maestro es verdad". La combinación de lo anterior podría conducir finalmente a que el imperativo "Haz tu tarea ahora" sea incluido entre los hechos bien establecidos, y eso, por la construcción de la máquina, significará que la tarea realmente empezará a ser hecha, pero el efecto es muy poco satisfactorio. El proceso de inferencia empleado por la máquina no tiene que satisfacer forzosamente a los lógicos más exigentes. Es posible, por ejemplo, que no haya jerarquía de tipos. Pero eso no tiene forzosamente que significar que surjan falacias de tipo con mayor probabilidad que la que nosotros tenemos de caer de acantilados sin muros. Los imperativos adecuados (expresados *dentro* de los sistemas y que no forman parte de las reglas *del* sistema), como por ejemplo: "No empleen una clase, a menos que sea una subclase de una clase mencionada por el maestro", pueden tener un efecto semejante a: "No se acerquen demasiado al borde".

Los imperativos que pueden ser obedecidos por una máquina que no tiene miembros serán forzosamente de carácter más bien intelectual, como en el ejemplo anterior (de hacer la tarea). Entre estos imperativos, son importantes los que regulan el orden en que las reglas del sistema lógico respectivo deben ser aplicadas. Porque en cada etapa en que empleamos un sistema lógico hay un número muy grande de medidas alternativas, de las cuales podemos aplicar cualquiera, siempre que obedezcamos las reglas del sistema lógico. Estas selecciones crean la diferencia entre un razonador brillante y uno trivial, y no la diferencia entre un razonador correcto y un razonador falaz. Las proposiciones que conducen a imperativos de este tipo podrían ser, por ejemplo: "Cuando se menciona a Sócrates, emplee el silogismo en Bárbara"; o: "Si se ha comprobado que un método es más rápido que otro, no emplee el método más lento". Algunos de ellos pueden ser "dados con autoridad", pero otros pueden ser producidos por la máquina misma, por ejemplo, por la inducción científica.

.....

NOTA 1 Confróntese una nueva edición de este argumento contra la inteligencia de las máquinas en Penrose, y la refutación de McCarthy, en el capítulo VI de esta antología. **Nota del editor**.

NOTA 2 Confróntese el ensayo de Dennett sobre la conciencia en el capítulo VI de esta antología. **Nota del editor**.

NOTA 3 Sin embargo, confróntese el artículo de Gutiérrez en el capítulo V de esta colección. **Nota del editor**.