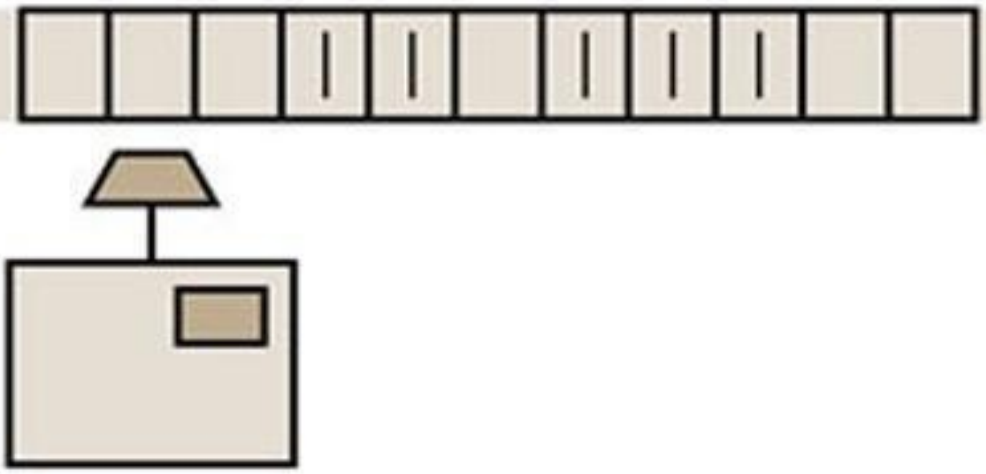
A.M. TURING

¿Puede pensar una máquina?

INTRODUCCIÓN DE MANUEL GARRIDO



En 1947 Alan M. Turing pronunció una conferencia ante un auditorio compuesto en su mayor parte por miembros del National Physical Laboratory de Londres en la que intentaba responder a la vieja y controvertida pregunta ¿Puede pensar una máquina?

Lo expuesto en ese acto apareció publicado tres años más tarde en Mind —una importante revista de filosofía británica— y es lo que ofrecemos aquí al lector en su traducción castellana. Este texto se convirtió enseguida en uno de los escritos fundacionales de la lógica informática y la inteligencia artificial, al presentar las líneas generales por las que debería discurrir una respuesta precisa y manejable (aunque no indiscutible) a la pregunta formulada.

Se trata del famoso Test de Turing, una prueba para decidir si una máquina es inteligente (o «piensa»). Para ello Turing diseñó un juego de imitación en el que participan una máquina y seres humanos; podemos decir que una máquina piensa si un ser humano que se comunica con la máquina y con otros seres humanos no logra distinguir cuando su interlocutor es una máquina y cuando un humano.

Una «máquina de Turing» como la que participa en el juego, es un dispositivo ideal de cálculo, capaz de resolver una función computable —una función cuya solución es susceptible de ser obtenida por un procedimiento mecánico —.

Pero lo más significativo es que Turing demostró que hay una máquina peculiar —la máquina universal de Turing— en la que se puede representar cualquier máquina que sea capaz de computar una función particular. De acuerdo con esto, una máquina universal de Turing sería una especie de sistema operativo en el que se implementan diferentes programas (máquinas de Turing especiales), un poco a la manera en que nos es familiar en los ordenadores personales. La denominada «metáfora del ordenador» como modelo capaz de simular la mente humana y, por ende, el pensar, tiene aquí su fuente.

Lectulandia

Alan M. Turing

¿Puede pensar una máquina?

Edición bilingüe CASTELLANO-INGLÉS

ePub r1.0
GusiX 24.02.14

Título original: *Computing machinery and intelligence*
Alan M. Turing, 1950
Traducción: Amador Antón y Manuel Garrido
Retoque de portada: GusiX

Editor digital: GusiX
ePub base r1.0

ESPAÑOL

1. El juego de imitación

Propongo que consideremos la siguiente pregunta: «¿Pueden pensar las máquinas?». Para empezar, definamos el significado de los términos «máquina» y «pensar», pero es una actitud peligrosa. Si hemos de llegar al significado de las palabras «máquina» y «pensar» a través de su utilización corriente, difícilmente escaparíamos a la conclusión de que hay que buscar el significado y la respuesta de la pregunta «¿Pueden pensar las máquinas?» mediante una encuesta tipo Gallup. Pero es absurdo. En lugar de intentar tal definición, sustituiremos la pregunta por otra estrechamente relacionada con ella y que se expresa con palabras relativamente inequívocas.

El problema en su nuevo planteamiento puede exponerse en términos de un juego que denominaremos «juego de imitación». Intervienen en él tres personas: un hombre (A), una mujer (B) y un preguntador (C), indistintamente de uno u otro sexo. El preguntador se sitúa en una habitación aparte y, para él, el juego consiste en determinar quién de los otros dos es el hombre y quién la mujer. Los conoce por la referencia X e Y, y al final del juego determina si «X es A e Y es B» o si «X es B e Y es A». El preguntador puede plantear a A y a B preguntas como éstas: «Por favor X, ¿podría decirme cuán largo es su pelo?»

Supongamos que X es realmente A, entonces es A quien contesta. El objetivo de A en el juego es lograr que C efectúe una identificación errónea, por lo que su respuesta podría ser: «Mi pelo es corto, escalonado, y los mechones más largos son de unos veinte centímetros».

Para que el preguntador no se guíe por el timbre de voz, las respuestas deben ir por escrito o, mejor aún, mecanografiadas. Lo ideal es disponer de un impresor telegráfico que comunique las dos habitaciones. Otro procedimiento consiste en que un intermediario repita pregunta y respuesta. El objeto del juego para el tercer jugador (B) es ayudar al preguntador. La mejor estrategia para la jugadora es probablemente responder la verdad, añadiendo quizás a sus respuestas cosas como ésta: «¡Soy la mujer, no le haga caso!», pero de nada sirve, ya que el hombre puede hacer observaciones similares.

Ahora planteemos la pregunta: «¿Qué sucede cuando una máquina sustituye a A en el juego?». ¿Se pronunciará el preguntador en este caso tan erróneamente como lo hace cuando en el juego participan un hombre y una mujer? Estas preguntas sustituyen a la original: «¿Pueden pensar las máquinas?»

2. Crítica del nuevo problema

Del mismo modo que preguntamos: «¿Cuál es la respuesta a este nuevo tipo de pregunta?», podemos preguntar: «¿Merece la pena resolver esta nueva pregunta?». Resolvamos esta última pregunta sin plantear más objeciones para cortar una regresión infinita.

El nuevo problema presenta la ventaja de que traza una línea definida entre las aptitudes físicas e intelectuales de una persona. Ningún ingeniero o químico puede atribuirse la capacidad de producir un material que no pueda distinguirse de la piel humana. Quizá sea posible algún día, pero, aun suponiendo la viabilidad de semejante invención, nos parece que de poco serviría tratar de hacer una «máquina pensante» más humana, forrándola con esa epidermis artificial. El modo en que hemos planteado el problema refleja el obstáculo que impide al preguntador ver o tocar a los otros concursantes, oír su voz. Otras ventajas del criterio propuesto pueden resumirse en un modelo de preguntas y respuestas. Por ejemplo:

P: Por favor, escriba un soneto sobre el tema del Cuarto Puente.
R: Hágame otra pregunta; la poesía no es mi fuerte.
P: Sume 34957 con 70764.
R: (Pausa de unos 30 segundos) 105621.
P: ¿Juega al ajedrez?
R: Sí.
P: Tengo el rey en la casilla 1R y ninguna otra pieza. Usted tiene sólo el Rey en la casilla 6R y la Dama en 1D. Le toca mover. ¿Qué juega?
R: (Pausa de unos 15 segundos) La Dama a D8, mate.

El método de preguntas y respuestas parece adecuado para introducir casi todos los campos de actividad humana que queramos. No vamos a sancionar a la máquina por su incapacidad para destacar en concursos de belleza, del mismo modo que no castigamos a una persona por perder una carrera en una competición aérea. Las condiciones del juego hacen irrelevantes esas torpezas. Los «testigos» pueden alardear, si lo creen conveniente, tanto como deseen con respecto a sus encantos, su fuerza o su heroísmo, pero el preguntador no puede exigir demostraciones fehacientes.

El juego quizá provoque críticas porque la máquina tiene demasiados factores en contra. Si una persona lo intentara haciéndose pasar por la máquina, sin duda haría un papel deplorable. Quedaría rápidamente eliminada por lentitud e inexactitud aritmética. ¿No harán las máquinas algo que permita la definición de pensamiento, pero que es muy distinto a lo que hace una persona? Se trata de una objeción de peso, pero cuando menos podemos decir que, dado que es posible construir una máquina

que realice satisfactoriamente el juego de imitación, la objeción no viene al caso.

Podría alegarse que la mejor estrategia en el «juego de imitación», para la máquina, es posiblemente algo distinto a la imitación de la conducta humana. Puede, pero yo no creo que esto influya demasiado. En cualquier caso, no nos proponemos aquí analizar la teoría del juego y supondremos que la mejor estrategia es tratar de dar las respuestas que una persona daría con toda naturalidad.

3. Las máquinas que intervienen en el juego

La cuestión que planteábamos en el apartado 1 carece de precisión si no especificamos qué entendemos por el término «máquina». Es lógico que deseemos que nuestras máquinas estén dotadas de cualquier tipo de ingeniería mecánica. Del mismo modo que aceptamos la posibilidad de que un ingeniero o un equipo de ingenieros construya una máquina que funcione, pero cuya modalidad operacional no pueden describir satisfactoriamente sus constructores porque se han servido de un método fundamentalmente experimental. Finalmente, excluiremos de la categoría de máquinas a las personas nacidas del modo habitual. Es difícil adaptar las definiciones de modo que cumplan estos tres requisitos. Se puede insistir, por ejemplo, en que el equipo de ingenieros sea de un solo sexo, lo cual no sería satisfactorio, ya que probablemente se puede crear un individuo completo a partir de una simple célula epidérmica de un hombre (pongamos por caso). Esto sería una proeza de biogenética merecedora de máxima admiración, pero no por ello la calificaríamos de «construcción de máquina pensante». Esto nos obliga a descartar el requisito de permitir cualquier tipo de técnica, y con mayor razón dado que el interés actual por las «máquinas pensantes» se ha suscitado gracias a un tipo particular de máquina, generalmente denominada «computadora electrónica» o «computadora digital». Con arreglo a esto, sólo permitiremos que tomen parte en el juego las computadoras digitales.

A primera vista esta limitación parece muy drástica, pero intentaré demostrar que no es así. Para ello es necesario un breve resumen sobre la naturaleza y las propiedades de estas computadoras. Podría también aducirse que esta identificación de las máquinas con las computadoras digitales, al igual que nuestro criterio sobre el término «pensar», son insatisfactorias si (en contra de lo que creo) resulta que las computadoras digitales son incapaces de hacer un buen papel en el juego.

Existen ya varías computadoras operacionales, y es lógico que se diga: «¿Por qué no realizar el experimento ahora mismo? No resultaría difícil cumplir los requisitos del juego. Se pueden utilizar varios preguntadores, compilando unas estadísticas para comprobar cuántas veces se produce la identificación correcta». La respuesta inmediata es que no se trata de plantearse si todas las computadoras digitales actuarán bien en el juego, ni de si las actuales computadoras actuarán bien, sino de si existen computadoras imaginables que actúen bien. Pero esto es sólo la respuesta inmediata, más adelante consideraremos la cuestión bajo otra perspectiva.

4. Computadoras digitales

Podemos explicar el concepto de computadoras digitales diciendo que son unas máquinas ideadas para realizar cualquier tipo de operación propia de un computador humano. El computador humano sigue unas reglas determinadas sin opción a desviarse de ellas bajo ningún concepto. Supongamos que esas reglas figuran en un libro que cambia cada vez que el computador acomete un nuevo trabajo. Dispone también de una cantidad ilimitada de papel para efectuar cálculos y hace las multiplicaciones y sumas pertinentes con una «máquina de bolsillo», pero esto no tiene importancia.

Si utilizamos como definición la anterior explicación, corremos el riesgo de caer en una argumentación circular. Para evitarlo, esbozaremos los medios con los que se logra el efecto deseado. Suele considerarse que una computadora digital consta de tres partes:

1. Almacenamiento
2. Unidad procesadora
3. Control

El almacenamiento es el acopio de información y corresponde al papel sobre el que se efectúa la computación humana, ya sea el papel en que la persona realiza los cálculos o aquél en el cual está impreso el libro de reglas. Del mismo modo que el computador humano efectúa sus cálculos con su cabeza, parte del almacenamiento corresponde a la memoria de la máquina.

La unidad procesadora es el sector que realiza las distintas operaciones de cálculo. La naturaleza de estas operaciones varía de una máquina a otra. Generalmente pueden efectuar operaciones bastante largas, tales como «Multiplicar 3540675445 por 7076345687», pero en algunas máquinas sólo pueden llevarse a cabo operaciones muy simples, tales como «Escribe 0».

Hemos mencionado que el «libro de reglas», de que se vale el computador, se sustituye en la máquina por una parte del almacenamiento. Esta se denomina «tabla de instrucciones». Corresponde al control comprobar que las instrucciones se sigan correctamente y en su debido orden. El control está construido de tal manera que es infalible.

La información almacenada suele estar dividida en paquetes de tamaño relativamente modesto. En una máquina concreta, por ejemplo, el paquete puede constar de diez dígitos decimales. Se asignan números a las partes del almacenamiento en que se guardan los diversos paquetes de información, con arreglo a una modalidad sistemática. Un ejemplo de instrucción corriente podría ser: «Suma la cifra almacenada en la posición 6809 a la situada en la 4302 y devuelve el

resultado de la última posición de almacenamiento». Ni que decir tiene que la operación no se desarrolla en la máquina expresada de este modo, sino que se lleva a cabo siguiendo una codificación como 6809430217. La cifra 17 indica cuál de las posibles operaciones hay que efectuar con las dos cifras. En cuyo caso la operación es la anteriormente descrita: «Suma la cifra…». Se advertirá que la instrucción consta de diez dígitos y, por lo tanto, constituye exactamente un paquete informativo. El control suele captar las instrucciones a seguir en el orden de posición en que están almacenadas, aunque a veces pueda surgir una instrucción como ésta: «Sigue ahora la instrucción almacenada en la posición 5606 y continúa», o bien: «Si la posición 4505 contiene 0, sigue la instrucción almacenada en 6707; en caso contrario continúa».

Las instrucciones de este tipo son muy importantes porque permiten la repetición de una secuencia de operaciones una y otra vez hasta que se cumple un determinado requisito, pero, al hacerlo, la máquina sigue en cada repetición, no nuevas instrucciones, sino las mismas indefinidamente. Recurramos a una analogía casera: supongamos que mamá desea que Tommy pase por el zapatero cada mañana camino del colegio para ver si han arreglado sus zapatos; puede decírselo cada mañana, o puede dejar una nota permanente en el vestíbulo para que el niño la vea al salir y recuerde que tiene que pasar por el zapatero, y luego, al volver, si trae los zapatos, rompa la nota. El lector debe aceptar como un hecho la construcción de computadoras digitales que, efectivamente, se han construido con arreglo a los principios expuestos y que realmente mimetizan con gran fidelidad los actos de un computador humano.

El libro de reglas que, según hemos señalado, utiliza el computador humano es, naturalmente, una ficción convencional. Los computadores humanos recuerdan en realidad lo que tienen que hacer. Si queremos hacer una máquina que mimetice el comportamiento de un computador humano en operaciones complicadas, hay que preguntarle a éste cómo lo hace y luego transferir la respuesta en forma de tabla de instrucciones. La elaboración de tablas de instrucciones suele denominarse «programación». La «programación de una máquina para que efectúe la operación A» significa insertar en la máquina la tabla de instrucción adecuada para que lleve a cabo A.

Una variante interesante de la idea de computadora digital es la «computadora digital con un elemento aleatorio». Estas máquinas disponen de instrucciones en las que interviene un dado o un proceso electrónico equivalente; una instrucción de este tipo puede ser, por ejemplo: «Arroja el dado y almacena la cifra resultante en 1000». A veces se las denomina máquinas de libre voluntad (aunque personalmente yo no utilice esta expresión). Normalmente no se puede determinar por simple observación de la máquina si ésta posee un elemento aleatorio, ya que se logra un efecto similar con dispositivos cuya elección depende de los dígitos de los decimales de π.

La mayoría de las computadoras digitales poseen un almacenamiento finito, aunque no existe dificultad teórica en la concepción de una computadora de almacenamiento ilimitado. Naturalmente, sólo podría utilizarse una parte finita de cada fase. De igual modo se habría podido construir una cantidad finita, pero cabe imaginar que sucesivamente fueran añadiéndose otras. Estas computadoras presentan especial interés teórico y las denominaremos computadoras de capacidad infinita.

El concepto de computadora digital es antiguo. Charles Babbage, profesor de matemáticas en la Universidad de Cambridge entre 1828 y 1839 concibió una a la que denominó Máquina Analítica, pero no la terminó. Aunque Babbage expuso los principios fundamentales, la máquina no representaba en aquella época gran interés. Su rapidez habría sido mucho mayor que la de un computador humano, pero unas 100 veces inferior a la de la máquina de Manchester, que a su vez es una de las máquinas modernas más lentas. El almacenamiento era puramente mecánico y se efectuaba por medio de ruedas y tarjetas.

El hecho de que la Máquina Analítica de Babbage estuviera concebida de forma totalmente mecánica nos ayudará a despejar cualquier superstición. Muchas veces se atribuye importancia al hecho de que las computadoras digitales modernas son eléctricas, igual que el sistema nervioso. Como la máquina de Babbage no era eléctrica, y como todas las computadoras digitales son en cierto modo equivalentes a ella, el empleo de la electricidad no es teóricamente relevante.

Siempre que se trata de señalización rápida interviene, claro, la electricidad. Por lo tanto, no es de extrañar que ésta se halle relacionada con ambos conceptos. En el sistema nervioso los fenómenos químicos son, cuando menos, tan importantes como los eléctricos. En ciertas computadoras el sistema de almacenamiento es fundamentalmente acústico. Por lo tanto, el empleo de la electricidad como propiedad no deja de ser una similitud muy superficial. Para establecer similitudes reales debemos más bien buscar analogías en el funcionamiento matemático.

5. Universalidad de las computadoras digitales

Podemos situar las computadoras digitales que hemos tratado en el apartado anterior dentro de la categoría de «máquinas de estado discreto». Estas son máquinas que pasan mediante saltos o clics súbitos de un estado bastante definido a otro. Se trata de estados lo bastante distintos para que no se dé la posibilidad de confusión entre ellos. Hablando en puridad no existen tales máquinas. En realidad, todo se mueve continuamente, pero podemos considerar positivamente muchos tipos de máquinas como de estado discreto. Por ejemplo, al referirnos a los interruptores de un sistema de iluminación, es una ficción convencional decir que cada uno de ellos debe hallarse totalmente conectado o desconectado. Pueden hallarse en posiciones intermedias, pero en la mayoría de los casos podemos descartarlas. Como ejemplo de máquina de estado discreto consideremos una rueda que recorra 120° por segundo, pero que se detiene al accionar una palanca externa; ésta, además, en determinada posición, enciende una luz. Podríamos definir esta máquina de forma abstracta del siguiente modo: El estado interno de la máquina (descrito por la posición de la rueda) puede ser q_1, q_2 o q_3. Hay una señal de entrada i_0 o i_1 (posición de la palanca). El estado interno en cualquier momento está determinado por el último estado, y la señal de entrada lo estará con arreglo a la tabla:

	Último Estado		
Entrada	q_1	q_2	q_3
i_1	q_2	q_3	q_1
i_0	q_1	q_2	q_3

Las señales de salida, única indicación visible externa del estado interno (la luz), nos las da la tabla

Estado	q_1	q_2	q_3
Salida	o_0	o_0	o_1

Es un ejemplo clásico de máquina de estado discreto. Este tipo de máquinas se describen por medio de las tablas indicadas, a condición de que posean únicamente un número finito de estados posibles.

Podría parecer que, dado el estado inicial de la máquina y la señal de entrada, siempre fuera posible predecir los estados futuros, pero es una reminiscencia de la perspectiva de Laplace, según la cual, a partir del estado completo del universo en un momento del tiempo, definido por las posiciones y velocidades de todas sus partículas, se pueden predecir los estados futuros. Sin embargo, la predicción que

estamos considerando es más próxima a la practicabilidad que la considerada por Laplace. El sistema del «universo como un todo» es de tal naturaleza que errores bastante pequeños en las condiciones iniciales pueden ejercer un efecto considerable en un momento futuro. El desplazamiento de un solo electrón en una billonésima de centímetro en un momento determinado puede ser la causa de que una persona muera aplastada por una avalancha un año más tarde o se libre de la catástrofe. Es una propiedad esencial de los sistemas mecánicos, que hemos denominado «máquinas de estado discreto», el que semejante fenómeno no se produzca. Incluso si consideramos las actuales máquinas físicas en lugar de las máquinas idealizadas, el conocimiento razonablemente exacto de su estado en determinado momento nos procura un conocimiento razonablemente exacto de cualquier serie de pasos ulteriores.

Como hemos dicho, las computadoras digitales pertenecen al grupo de máquinas de estado discreto. Pero el número de estados que pueden adoptar este tipo de máquinas suele ser enormemente elevado. Por ejemplo, para la máquina que actualmente funciona en Manchester, la cifra aproximada sería de 2^{165000}, es decir de 10^{50000} aproximadamente. Compárese esto con el citado ejemplo de la rueda que tenía tres estados. Se comprende sin dificultad por qué es tan elevado el número de estados. La computadora posee un almacenamiento correspondiente al papel que utiliza un computador humano. En este almacenamiento puede escribirse cualquiera de las combinaciones de símbolos que figurasen en el papel. Para simplificar, supongamos que sólo utilizamos como símbolos los dígitos del 0 al 9. No tomaremos en cuenta las variaciones de los signos manuscritos. Supongamos que la computadora dispone de 100 hojas de papel de 50 líneas cada una, con espacio para 30 dígitos. El número de estados será $10^{10\times50\times30}$, es decir, 10^{150000}. Esto equivale aproximadamente al número de estados de tres máquinas de Manchester juntas. El logaritmo con base dos del número de estados es en realidad lo que se denomina «capacidad de almacenamiento» de la máquina. Por lo tanto, la máquina de Manchester posee una capacidad de almacenamiento aproximada de 165000, y la máquina con rueda del ejemplo mencionado, de aproximadamente 1,6. Si juntamos dos máquinas, habrá que sumar sus capacidades para saber la capacidad de la máquina resultante. Esto nos permite afirmar que «la máquina de Manchester contiene 64 pistas magnéticas, cada una de ellas con capacidad para 2560, ocho tubos electrónicos con capacidad de 1280. El almacenamiento diverso equivale aproximadamente a 300, lo que da un total de 174380».

Disponiendo de la tabla correspondiente a una máquina de estado discreto se puede predecir lo que hará, y nada nos impide efectuar este cálculo con una computadora digital. A condición de que lo efectúe con suficiente rapidez, la computadora digital puede mimetizar el comportamiento de cualquier máquina de estado discreto. Entonces, se podría jugar con esa máquina (en el papel B) al juego de

imitación y con la computadora digital mimetizante (en el papel de A), y el interrogador no sabría diferenciarlas. Naturalmente, la computadora digital debe poseer una capacidad de almacenamiento adecuada y funcionar a suficiente velocidad. Además, habrá que programarla expresamente para cada nueva máquina que se desee imitar.

Esta propiedad esencial de las computadoras digitales, por la que pueden imitar a cualquier máquina de estado discreto, se define diciendo que son máquinas *universales*. La existencia de máquinas con esta propiedad encierra la importante consecuencia de que, consideraciones de rapidez aparte, no hay necesidad de diseñar diversas máquinas nuevas para que realicen los correspondientes nuevos procesos de computación. Todos pueden efectuarse con una sola computadora digital, convenientemente programada en cada caso. En consecuencia, como veremos, todas las computadoras digitales de este tipo son equivalentes en un sentido.

Ahora consideraremos la cuestión mencionada al final del apartado 3. Habíamos sugerido sustituir la pregunta «¿Pueden pensar las máquinas?» por la de «¿Existen computadoras digitales imaginables que jueguen bien al juego de imitación?». Si se desea, puede generalizarse más superficialmente esta pregunta: «¿Hay máquinas de estado discreto que hagan un buen juego?». Pero, dada la propiedad universal, vemos que ambas preguntas equivalen a: «Supongamos una determinada computadora digital C. ¿Es cierto que, modificando esta computadora para que tenga un almacenamiento adecuado y dotándola de un programa apropiado, podemos conseguir que C desempeñe eficazmente el papel de A en el juego de imitación y el papel de B lo haga un hombre?»

6. Opiniones contrapuestas sobre la cuestión principal

Consideremos ahora que hemos despejado el terreno y podemos ya pasar al debate de la pregunta «¿Pueden pensar las máquinas?» y de su variante, expuesta al final del apartado anterior. No podemos descartar totalmente la forma original del problema, ya que habrá diversidad de opiniones con respecto a la pertinencia de la sustitución y no podemos por menos que atender lo que se diga sobre el asunto.

Simplificaré las cosas para el lector si, en primer lugar, explico mi propia opinión sobre el tema. Consideremos primero la forma más exacta de la pregunta. Personalmente creo que, dentro de unos cincuenta años, se podrá perfectamente programar computadoras con una capacidad de almacenamiento aproximada de 10^9 para hacerlas jugar tan bien al juego de imitación que un preguntador corriente no dispondrá de más del 70 por ciento de las posibilidades para efectuar una identificación correcta a los cinco minutos de plantear las preguntas. Me parece que la pregunta original, «¿Pueden pensar las máquinas?», no merece discusión por carecer de sentido. No obstante, creo que, a finales del siglo, el sentido de las palabras y la opinión profesional habrán cambiado tanto que podrá hablarse de máquinas pensantes sin levantar controversias. Creo además que de nada sirve ocultar las ideas. La opinión tan generalizada de que los científicos proceden siempre de un hecho bien demostrado a otro hecho bien demostrado, y nunca se dejan influir por una conjetura no probada, es bastante errónea. A condición de que quede bien claro qué son hechos probados y qué son conjeturas, no existe ningún peligro. Las conjeturas son de suma importancia, porque sugieren posibles vías de investigación.

Ahora consideraré opiniones contrarias a la mía:

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 no a ningún animal ni máquina. Por lo tanto, ni los animales ni las máquinas pueden pensar.

Personalmente son ideas que rechazo totalmente, pero intentaré refutarlas en términos teológicos. La argumentación resultaría más convincente si se clasificara a los animales con el hombre, ya que existe mucha más diferencia, para mí, entre lo genuinamente animado y lo inanimado que entre el hombre y los animales. El carácter arbitrario de la opinión ortodoxa se evidencia aún más si tenemos en cuenta la opinión de los creyentes de otras religiones. ¿Cómo ve el cristianismo el dogma musulmán según el cual la mujer no tiene alma? Pero dejemos esto y volvamos a la cuestión principal. Creo que el citado argumento implica una grave restricción de la

omnipotencia del Todopoderoso. Se admite así que hay cosas de las que Él es incapaz, como es hacer que uno sea igual a dos, pero ¿dudaremos de su libertad para insuflar alma a un elefante, si a bien lo tiene? Cabe esperar que únicamente ejerciese tal poder en conjunción con una mutación que dotase al elefante de un cerebro mejorado que respondiera a las necesidades de esa alma. Podemos argüir exactamente lo mismo en el caso de las máquinas. Puede parecer distinto por ser más difícil de «tragar», pero esto únicamente significa que pensamos que es menos verosímil que El considere adecuadas las circunstancias para dotarlas de alma. Las circunstancias en cuestión se discuten en el resto de este trabajo. Al intentar construir este tipo de máquinas no estamos usurpando irreverentemente Su poder de crear almas, igual que no lo hacemos al procrear niños; en realidad, en ambos casos somos instrumentos de Su voluntad al procurar moradas para las almas que Él crea.

Pero todo esto es mera especulación. No me impresionan mucho los argumentos teológicos, aunque se utilicen como apoyo. A lo largo de la historia se ha comprobado cuánto dejan que desear. En tiempos de Galileo se argumentaba que las Sagradas Escrituras decían: «Y el sol se detuvo… y no fue hacia el ocaso durante casi un día» (Josué x.13) y que: «El creó los fundamentos de la Tierra para que no se moviera» (Salmo cv. 5) como refutación convincente de la teoría copernicana. Con los conocimientos actuales estos argumentos resultan fútiles, pero en una época de escasos conocimientos científicos causaban muy distinta impresión.

2. La objeción del «avestruz»

«Las consecuencias de que las máquinas piensen serían horribles. Creamos y esperemos que no sea posible».

Este argumento rara vez se expone de forma tan abierta, pero afecta a la mayoría de quienes reflexionamos sobre ello. Nos gusta creer que el hombre es en algún modo superior al resto de la creación, y tanto mejor si podemos demostrar que es necesariamente superior, pues entonces no existe peligro de que pierda su posición dominante. La popularidad del argumento teológico está claramente vinculada a esta idea y cuenta con muchos adeptos entre los intelectuales, pues éstos aprecian más que otras personas el poder del pensamiento y se muestran más inclinados a basar su convencimiento de la superioridad del hombre en este poder.

No creo que este argumento sea lo bastante fundado para molestarme en refutarlo. Tal vez sea mejor consolarse, buscándolo quizás en la transmigración de las almas.

3. La objeción matemática

Pueden citarse toda una serie de resultados de la lógica matemática para demostrar que hay limitaciones en el poder de las máquinas de estado discreto. El más conocido es el denominado teorema de Gödel, que demuestra que en cualquier sistema lógico lo bastante potente pueden formularse afirmaciones que no pueden demostrarse ni refutarse dentro del sistema, salvo en caso de que posiblemente tal sistema sea incoherente. En ciertos aspectos similares hay otros resultados expuestos por Church, Kleene, Rosser y Turing. La tesis de este último autor es la que merece mayor consideración en este caso, por referirse específicamente a las máquinas, mientras que las de los otros sólo son utilizables en tanto que argumentos relativamente indirectos: si, por ejemplo, recurrimos al teorema de Gödel, hace falta a la vez disponer de medios para describir los sistemas lógicos en términos de máquinas y las máquinas en términos de sistemas lógicos. El resultados en cuestión se refiere a un tipo de máquina que es fundamentalmente una computadora digital con capacidad infinita, y postula que hay ciertas cosas que esa máquina no puede efectuar. Si se la equipa para dar respuesta a preguntas como en el juego de imitación, habrá preguntas que contestará mal o no podrá contestar por mucho tiempo que se le conceda. Naturalmente, puede haber muchas preguntas de esta clase, y preguntas que no pueda contestar satisfactoriamente una máquina las contestará adecuadamente otra. Desde luego estamos por ahora en la suposición de que estas preguntas son de tal índole que la respuesta es «Sí» o «No», y no preguntas del tipo «¿Qué opinas sobre Picasso?». Las preguntas que sabemos que la máquina no contesta son de esta clase: «Supongamos una máquina con las siguientes características… ¿contestará esta máquina “Sí” a cualquier pregunta?». Los puntos suspensivos se sustituyen por la descripción de una máquina modelo estándar, que podría ser como la que se cita en el apartado 5. Si la máquina descrita guarda cierta relación comparativamente simple con la máquina a que se está interrogando, puede demostrarse que la respuesta es incorrecta o no se va a producir. Este es el resultado matemático: se arguye que demuestra una incapacidad por parte de las máquinas a la que no está expuesto el intelecto humano.

La respuesta taxativa a este razonamiento es que, aunque está demostrado que existen limitaciones en la capacidad de cualquier máquina, sólo se ha afirmado, sin prueba alguna, que tales limitaciones no son aplicables al intelecto humano. Sin embargo, yo no creo que esta posibilidad pueda rechazarse tan alegremente. Cuando se plantea a una de estas máquinas la pregunta crítica adecuada y nos da una respuesta concreta, sabemos que la respuesta es incorrecta y esto nos da cierta sensación de superioridad. ¿Es una sensación ilusoria? Sin duda es lo bastante legítima, pero yo no creo que haya que atribuirle demasiada importancia. También nosotros en muchas ocasiones respondemos erróneamente a preguntas, lo cual no justifica esa enorme sensación de halago al ver que las máquinas fallan. Además, sólo

podemos sentir en este caso nuestra superioridad en relación con la máquina concreta, objeto de nuestra frágil victoria. No es un triunfo simultáneo frente a *todas* las máquinas. En resumen, habrá hombres más listos que cualquier máquina, pero también otras máquinas más listas, y así sucesivamente.

Los partidarios del argumento matemático aceptarán en su mayoría —creo yo— que el juego de imitación es una buena base para la discusión. A los partidarios de las dos primeras objeciones seguramente no les interesará ningún razonamiento.

4. El argumento de la conciencia

Este argumento está perfectamente expresado en un discurso conmemorativo del profesor Jefferson, en 1949, del que cito: «Hasta que una máquina sea capaz de escribir un soneto o de componer un concierto, porque tenga la facultad de reflexionar y sea capaz de sentir, y no por la combinación aleatoria de símbolos, no podremos admitir que esa máquina sea igual al cerebro, en el sentido de que no sólo los escriba, sino que sepa que los ha escrito. Ningún mecanismo (y no hablo de una señal artificial, invención simplona) puede sentir placer por sus logros, pena cuando se funden sus válvulas, regocijo por los halagos, depresión por sus errores, atracción sexual, enfado o decepción cuando no consigue lo que quiere».

Este argumento parece ser la negación de la validez de nuestro test. Según la modalidad más extremada de este tipo de planteamiento, la única manera de asegurarse de que una máquina piensa es ser la máquina y sentir el propio pensamiento. Sólo entonces pueden exponerse tales sentimientos a todo el mundo, pero tampoco está justificado que a nadie le importen. Según este planteamiento, también la única manera de saber que una persona piensa es ser esa persona concreta. De hecho, es un punto de vista solipsista. Puede que sea el punto de vista más lógico, pero dificulta la comunicación de ideas. A puede sentirse inclinado a creer «A piensa pero B no», mientras que B creerá que «B piensa pero A no». En lugar de discutir indefinidamente este punto, mejor es adscribirse al cortés convencionalismo de que todos piensan.

Estoy convencido de que el profesor Jefferson no desea adoptar el punto de vista extremo y solipsista. Probablemente se halle dispuesto a aceptar como prueba el juego de imitación. El test (omitiendo el jugador B) suele usarse en la práctica bajo la denominación de *examen oral* para descubrir si el candidato entiende de verdad algo o lo «ha aprendido como un papagayo». Escuchemos un extracto de uno de esos *exámenes orales*:

Examinador: En el primer verso de su soneto, que dice «¿Te compararía con un día de verano?», ¿no sería igual, o mejor, «un día de primavera»?

Examinado: No rimaría.

Examinador: ¿Y «un día de invierno»? Rima perfectamente.

Examinado: Sí, pero a nadie le gusta que le comparen con un día de invierno.

Examinador: ¿Diría usted que Mr. Pickwick le recuerda la Navidad?

Examinado: En cierto modo.

Examinador: Pues Navidad es un día de invierno, y no creo que a Mr. Pickwick le molestara la comparación.

Examinado: Creo que bromea usted. Por día de invierno se entiende un día de invierno genuino y no uno especial como el de Navidad.

Y así sucesivamente. ¿Qué diría el profesor Jefferson si la máquina escritora de sonetos fuera capaz de contestar así en el examen oral? No sé si la consideraría accionada por «una simple señal artificial» al dar tales respuestas, pero, si las respuestas fueran tan adecuadas y coherentes como en el párrafo anterior, no creo que las calificara de «invención simplona». Yo creo que con esta expresión se intenta definir dispositivos tales como la inclusión en la máquina de un disco de alguien que lee un soneto, dotado del correspondiente relé que lo conecte de vez en cuando.

En resumen, creo que a la mayoría de los partidarios del argumento de la conciencia se les podría convencer de que lo abandonaran en lugar de forzarles a la actitud solipsista. Entonces, probablemente se inclinasen a aceptar la prueba.

No quisiera dar la impresión de que creo que no existe misterio en lo que se refiere a la conciencia. Existe, por ejemplo, algo así como una paradoja en relación con su localización. Pero no creo que haya que solucionar necesariamente ese misterio para responder a la cuestión que nos ocupa en este trabajo.

5. Argumentos de incapacidades diversas

Estos argumentos responden al esquema: «Te aseguro que pueden hacerse máquinas que realicen todo lo que has dicho, pero es imposible construir una máquina que haga X», y se citan al respecto diversas X. A continuación expongo una selección:

«Ser amable, ingeniosa, hermosa, amistosa», «poseer iniciativa, tener sentido del humor, distinguir entre lo bueno y lo malo, cometer faltas», «enamorarse, apreciar las fresas y los helados», «enamorar a alguien, aprender por la experiencia», «utilizar adecuadamente las palabras, ser objeto de su propio pensamiento», «tener un comportamiento tan versátil como una persona, hacer algo auténticamente nuevo».

Generalmente estas afirmaciones no se apoyan en razonamientos y, personalmente, creo que en esencia se basan en el principio de la inducción científica. Una persona ve miles de máquinas durante su vida y, por lo que ve de ellas, extrae

una serie de conclusiones generales. Son feas y cada una de ellas está ideada para una tarea concreta; cuando se desea que ejecuten varias funciones, son inservibles, su variedad de comportamiento es muy limitada, etc., etc. En consecuencia, concluye que ésas son las características de las máquinas en general. Muchas de estas limitaciones se asocian a la escasa capacidad de almacenamiento de la mayoría de las máquinas (supongo que, en el concepto de capacidad de almacenamiento, se incluyen en cierto modo a las máquinas distintas a las de estado discreto. No importa la definición exacta, ya que no aspiramos a una exactitud matemática en esta discusión). Hace unos años, cuando aún se hablaba poco de computadoras digitales, era de esperar que su mención suscitara incredibilidad cuando se hablaba de sus propiedades sin explicar su construcción. Supongo que era también debido a la aplicación del principio de inducción científica. Naturalmente esta clase de aplicación del principio suele ser inconsciente. Cuando un niño que ha sufrido una quemadura teme al fuego y demuestra que lo teme evitándolo, decimos que está aplicando la inducción científica. (Naturalmente, puedo también describir su comportamiento de muchas otras maneras). Los trabajos y las costumbres humanos no parecen constituir un material muy adecuado para la aplicación de la inducción científica. Habría que investigar una gran magnitud espacio-temporal para obtener resultados fiables, pues, si no, creeremos (como la inmensa mayoría de los niños ingleses) que todo el mundo habla inglés y que es una tontería aprender francés.

Sin embargo, conviene hacer algunas observaciones respecto de las múltiples incapacidades que hemos citado. La incapacidad para apreciar las fresas y los helados le habrá parecido al lector una futilidad. Puede que se construya una máquina que aprecie esos manjares, pero sería una imbecilidad intentarlo. Lo importante respecto de esta incapacidad es que está destinada a aumentar el número de incapacidades, por ejemplo, el mismo tipo de dificultad de comunicación amistosa que se produce entre el hombre y la máquina también se da entre un hombre blanco y otro hombre blanco, o entre un hombre negro y otro hombre negro.

Afirmar que las «máquinas no cometen errores» parece curioso. Se siente uno inclinado a replicar: «¿Y son por eso peores?», pero adoptemos una actitud más simpática y tratemos de comprender qué es lo que significa. Creo que esta crítica puede explicarse en términos del juego de imitación. Se afirma que al preguntador le basta, para distinguir una máquina del hombre, plantear una serie de problemas aritméticos. La máquina queda desenmascarada por su tremenda exactitud. Así de sencillo, pero la máquina (programada para jugar el juego) no tratará de dar las respuestas *correctas* a los problemas aritméticos e introducirá deliberadamente errores de modo calculado para confundir al preguntador. Una avería mecánica se percibirá probablemente al darse una decisión inadecuada respecto del tipo de error aritmético a efectuar. Incluso esta interpretación crítica no es lo bastante simpática,

pero no disponemos de espacio para extendernos más. A mí me parece que la crítica se fundamenta en una confusión de dos tipos de error. Podemos denominarlos «errores de funcionamiento» y «errores de conclusión». Los errores de funcionamiento los causa un efecto mecánico o eléctrico que obliga a la máquina a comportarse de modo distinto a como está diseñada. En las discusiones filosóficas se ignora la posibilidad de tales errores y se habla de «máquinas abstractas». Estas máquinas abstractas son ficciones matemáticas más que objetos físicos. Son, por definición, incapaces de errores de funcionamiento. En este sentido podemos afirmar con certeza que «las máquinas no cometen errores». Los errores de conclusión sólo pueden producirse cuando se atribuye un significado a las señales de salida de la máquina. La máquina puede, por ejemplo, imprimir ecuaciones matemáticas, o frases en inglés. Cuando escribe una oración incorrecta, decimos que ha cometido un error de conclusión. Evidentemente no existe motivo para decir que una máquina no puede cometer este tipo de error. Puede que se limite a escribir sin parar «0 = 1». Adoptando un ejemplo menos peyorativo, digamos que, al estar dotada de un método para extraer conclusiones por inducción científica, es presumible que semejante método conduzca a veces a resultados erróneos.

A la afirmación de que una máquina no puede ser objeto de su propio pensamiento sólo puede contestarse si se demuestra que la máquina posee algún pensamiento referido a *algún* tema. No obstante, «el tema de las operaciones de una máquina» parece significar algo, al menos para quienes trabajan con ella. Si, por ejemplo, la máquina trata de hallar la solución a la ecuación $x^2 - 40x - 11 = 0$, uno no puede resistir la tentación de calificar esta ecuación de objeto parcial del tema de la máquina en ese momento. En este aspecto no cabe duda de que una máquina es su propio objeto, ya que se la puede utilizar para que contribuya a la confección de su propio programa, o para predecir el efecto de alteraciones en su propia estructura. Observando los resultados de su propio comportamiento, es capaz de modificar sus programas para efectuar determinada tarea con mayor eficacia. Son posibilidades de un futuro no muy lejano, no sueños utópicos.

La crítica de que una máquina no puede tener versatilidad de comportamiento es sólo una manera de decir que no puede tener una gran capacidad de almacenamiento. Hasta hace relativamente poco tiempo una simple capacidad de mil dígitos era algo extraordinario.

Las críticas que estamos considerando suelen ser variantes enmascaradas del argumento de la conciencia. Generalmente, si uno sostiene que una máquina *puede* hacer una de esas cosas y describe la clase de método del que puede servirse, no se logra impresionar a los detractores, pues piensan que el método (sea el que fuere, por ser mecánico necesariamente) es algo vil. Cotéjese el paréntesis del párrafo de Jefferson citado anteriormente.

6. Objeción de lady Lovelace

La información más pormenorizada sobre la máquina analítica de Babbage figura en un informe de lady Lovelace. En él se afirma: «La Máquina Analítica no pretende *crear* nada. Puede realizar *lo que nosotros sepamos mandarle*» (en cursiva en el informe original). Es Hartree quien cita este párrafo, y añade: «Esto no implica que sea imposible construir equipo electrónico que "piense por sí solo", o en el que, en términos biológicos, no se pueda implantar un reflejo condicionado que sirva de base al "aprendizaje". Si es o no posible en principio, es una cuestión apasionante y estimulante, esbozada en algunos de los últimos avances tecnológicos. Pero no parecía que las máquinas construidas en aquella época tuvieran tal propiedad».

Coincido totalmente con Hartree al respecto. Adviértase que él no afirma que la máquina en cuestión no posea la propiedad, sino que a lady Lovelace no le constaba que la tuviera. Es muy posible que las máquinas en cuestión tuvieran en cierto modo esa propiedad. Supongamos que una máquina de estado discreto tiene esa propiedad. La Máquina Analítica era una computadora digital universal, de forma que, si su capacidad de almacenamiento y su velocidad eran adecuados, con un programa idóneo se la podría inducir a mimetizar la propia máquina. Probablemente este razonamiento no se le ocurrió a la condesa ni al propio Babbage. En cualquier caso, ellos no tenían por qué reivindicar todo lo reivindicable.

Volveremos a hablar del tema en el apartado de máquinas que aprenden.

Una variante a la objeción de lady Lovelace afirma que las máquinas «nunca hacen nada nuevo». Podemos parangonar tal afirmación al refrán: «No hay nada nuevo bajo el sol». ¿Quién puede tener el firme convencimiento de que el «trabajo original» que se acaba de realizar no es sino el desarrollo de la simiente que ha dejado en él el aprendizaje, o la consecuencia de atenerse a consabidos principios generales? Otra variante mejor de esta objeción es la de que la máquina nunca «puede sorprendernos». Es un desplante más directo, por lo que respondemos directamente. Las máquinas me sorprenden muy a menudo. Fundamentalmente porque no calculo lo suficiente para figurarme lo que van a hacer, o, más bien, porque, aunque calculo, lo hago de forma precipitada, descuidada y corriendo riesgos, y me digo: «Supongo que el voltaje es aquí el mismo que allí; bueno, supongamos que es el mismo». Naturalmente, muchas veces me equivoco, el resultado me sorprende, aunque, una vez finalizado el experimento, me olvido de mis falsas suposiciones. Con esta confesión me expongo a sermones sobre mis malas costumbres, pero no empaño mi sinceridad al dar fe de las sorpresas que experimento.

No pretendo con esta réplica silenciar la crítica. Probablemente puede deducirse que tales sorpresas se deben a algún acto creativo mental por mi parte, y nada dicen a favor de la máquina. Esto nos obliga a volver al argumento de la conciencia, muy

lejos de la idea de sorpresa. Es un tipo de argumentación muy similar, pero quizá valga la pena señalar que la apreciación de algo como sorprendente requiere igual «acto mental creativo», independientemente de que la sorpresa la cause una persona, un libro, una máquina o lo que sea.

La opinión de que las máquinas no pueden producir sorpresa se basa, creo yo, en el sofisma en el que suelen incurrir particularmente filósofos y matemáticos: la asunción de que, cuando a la mente se le presenta un hecho, todas las consecuencias del mismo la invaden con él simultáneamente. Es una asunción muy útil en muchas circunstancias, pero se olvida con harta facilidad de que es falsa. Una consecuencia natural de asumirla como cierta es que se da por sentado que no hay mérito en la simple elucidación de consecuencias a partir de datos y principios generales.

7. Argumento de la continuidad del sistema nervioso

Desde luego el sistema nervioso no es una máquina de estado discreto. Un pequeño error de información sobre la magnitud de un impulso nervioso aferente en una neurona puede modificar considerablemente la magnitud del impulso de salida. Puede argüirse que, precisamente por eso, no cabe posibilidad de mimetizar el comportamiento del sistema nervioso mediante un sistema de estado discreto. Cierto es que una máquina de estado discreto es distinta a una máquina continua, pero, si nos ceñimos a las condiciones del juego de imitación, el preguntador no gana nada con esa diferencia. Podemos aclarar la situación si consideramos cualquier otra máquina continua más sencilla. Un analizador diferencial, por ejemplo. (Un analizador diferencial es un tipo de máquina de estado no discreto que se emplea para cierta clase de cálculos). Algunos dan la respuesta impresa, por lo que son adecuados para intervenir en el juego. Una computadora digital no puede predecir exactamente las respuestas que da a un problema el analizador diferencial, pero sí puede dar la respuesta correcta. Por ejemplo, si se pregunta el valor de π (3,1416 aproximadamente), es razonable elegir 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 (pongamos por caso). En tales circunstancias resultará muy difícil para el preguntador distinguir al analizador diferencial de la computadora digital.

8. El argumento de la informalidad de comportamiento

No se puede elaborar un conjunto de reglas para describir lo que una persona hace en todas las circunstancias concebibles. Puede establecerse la regla de que, por ejemplo, hay que detenerse al ver un semáforo rojo y continuar si se ve uno verde,

pero ¿qué sucede si, por un error, se iluminan los dos a la vez? Quizá la persona decida que es mejor detenerse. Pero por esta decisión pueden surgir ulteriormente dificultades. Intentar sentar reglas de conducta que cubran cualquier eventualidad, hasta las resultantes de las luces de tráfico, parece imposible. Estoy de acuerdo con esto.

A partir de ello se arguye que no podemos ser máquinas. Trataré de exponer el argumento, pero temo no hacerle debidamente justicia. Al parecer, se desarrolla de este modo: «Si cada persona posee un conjunto fijo de reglas de conducta por las que rige su vida, no sería más que una máquina; pero no hay tales reglas. Por lo tanto, las personas no pueden ser máquinas». Es deslumbrante el injusto medio. No creo que el argumento se plantee casi nunca así, pero estoy convencido de que constituye la base de la argumentación. Sin embargo, puede darse cierta confusión entre «reglas de conducta» y «leyes de comportamiento» para oscurecer la conclusión. Por «reglas de conducta» entiendo preceptos tales como «Pare si ve luces rojas» que uno puede cumplir conscientemente. Por «leyes de comportamiento» entiendo leyes naturales aplicables al cuerpo humano, tales como «si le pellizcas, chilla». Si sustituimos «leyes de comportamiento que regulan su vida» por «leyes de conducta por las que rige su vida», el injusto medio deja de ser insuperable en el argumento en cuestión, pues creemos que no sólo es cierto que estar regulado por leyes de comportamiento implica ser una especie de máquina (aunque no necesariamente una máquina de estado discreto), sino que, a la inversa, ser tal máquina implica estar regulado por tales leyes. Sin embargo, no podemos convencernos tan fácilmente de la ausencia total de leyes de comportamiento como de la ausencia absoluta de leyes de conducta. El único modo de descubrir tales leyes consiste en la observación científica, y no conocemos circunstancias en las que pueda decirse: «Ya hemos buscado bastante. No existen tales leyes».

Podemos demostrar más categóricamente que semejante afirmación es injustificada. Supongamos que fuera posible con absoluta seguridad descubrir esas leyes, si existiesen. Entonces, dada una máquina de estado discreto, no cabe duda de que podría descubrirse, mediante la observación suficiente para predecirlas, su comportamiento futuro, y eso dentro de un tiempo razonable, digamos mil años. Pero no parece ser el caso. He elaborado en la computadora de Manchester un pequeño programa con tan sólo 1000 unidades de almacenamiento, merced al cual, si se entrega a la máquina una cifra de dieciséis guarismos, responde con otra de igual magnitud en dos segundos. Desafío a cualquiera a que descubra en esas respuestas suficientes datos sobre el programa para ser capaz de predecir cualquier respuesta a valores no probados.

9. El argumento de la percepción extra-sensorial

Supongo que el lector está al corriente de la idea de percepción extra-sensorial y del significado de sus cuatro variantes: telepatía, clarividencia, precognición y psicocinesis. Estos extraños fenómenos parecen refutar todas las ideas cinéticas habituales, ¡Cuánto nos gustaría desacreditarlos! Pero lamentablemente la evidencia estadística, al menos en el caso de la telepatía, es abrumadora. Resulta difícil para cualquiera reajustar sus propias ideas para dar cabida a estos hechos singulares, pero, una vez admitidos, no parece que cueste mucho creer en fantasmas y espíritus. Lo primero que se nos ocurre es la idea de que nuestros cuerpos se mueven de modo simple con arreglo a las leyes físicas conocidas, junto a otras no descubiertas pero bastante parecidas.

Para mí es un argumento de bastante peso. Podría argüirse que muchas teorías científicas siguen siendo válidas en la práctica, a pesar de que contradigan la percepción extra-sensorial, y que puede prescindirse perfectamente de ella, pero no deja de ser un conformismo fácil; precisamente es muy de temer que no sea el pensamiento el tipo de fenómeno en el que la percepción extra-sensorial sea particularmente relevante.

Un argumento más específico basado en la percepción extra-sensorial sería el siguiente: «Juguemos al juego de imitación, teniendo por testigo a una persona que sea buena receptora telepática y a una computadora digital. El preguntador puede plantear preguntas de este tipo: “¿A qué palo pertenece la carta que tengo en mi mano derecha?”. La persona, mediante telepatía o clarividencia, da la respuesta correcta 130 veces sobre 400 cartas. La máquina sólo puede adivinar al azar y tal vez acierte 104 veces, y así el preguntador efectúa la identificación correcta». Esta es una interesante posibilidad. Supongamos que la computadora digital cuenta con un generador numérico aleatorio, es natural que lo utilice para dar la respuesta. Pero el generador numérico aleatorio está sujeto al poder psicocinético del preguntador y quizás esta psicocinesis sea la causa de que la máquina acierte más veces de las que cabría esperarse de un cálculo de probabilidades, por lo que el preguntador seguiría siendo incapaz de efectuar la identificación correcta. Por otra parte, puede ser capaz de acertar sin plantear preguntas, gracias a la clarividencia. Con la percepción extra-sensorial puede suceder cualquier cosa.

Si admitimos la telepatía, habrá que depurar la prueba. Puede considerarse la situación similar a la que se produce si el preguntador hablara consigo mismo y uno de los participantes estuviera escuchando con el oído en la pared. Situando a los participantes en una «habitación a prueba de telepatía», se restablecerían las condiciones.

7. Máquinas que aprenden

Habrá comprobado el lector que no dispongo de argumento positivo alguno lo bastante convincente para apoyar mi tesis. Si lo tuviera, no me habría tomado tanta molestia en exponer detalladamente las falacias de las tesis contrarias. Ahora expondré la evidencia en favor de mi punto de vista.

Volvamos brevemente a la objeción de lady Lovelace, quien afirmaba que la máquina sólo puede hacer lo que nosotros le mandemos. Podríamos decir que una persona puede «inyectar» una idea en una máquina y que ésta respondería hasta cierto límite, quedándose quieta a continuación, como la cuerda de un piano percutida por un martillo. Otro símil podría ser una pila atómica de tamaño inferior al crítico: una idea inyectada correspondería a un neutrón que penetra desde fuera en la pila. Cada uno de estos neutrones provoca una determinada alteración que acaba por disiparse. Sin embargo, si aumentamos suficientemente el tamaño de la pila, la alteración causada por el neutrón incluso irá en aumento hasta la completa destrucción de la pila. ¿Existe un fenómeno equivalente para las mentes, y se da también en el caso de las máquinas? En el caso de la mente humana parece haberlo. La mayoría de los cerebros parecen ser «subcríticos», es decir, que corresponden en esta analogía a pilas de tamaño subcrítico. Una idea presentada a este tipo de mente, no inducirá generalmente más que una idea por respuesta. Una reducidísima proporción de cerebros son supercríticos. En ellos una idea da origen a toda una «teoría» formada por ideas secundarias, terciarias y de todo orden. Las mentes animales parecen decididamente ser subcríticas. Siguiendo la analogía, nos preguntamos: «¿Se puede hacer que una máquina sea supercrítica?»

La analogía de la «piel de cebolla» también es válida. Si consideramos las funciones de la mente o del cerebro, observamos determinadas operaciones explicables en términos puramente mecánicos. Lo que decimos no es aplicable a la auténtica mente: es una especie de piel que hay que quitar si queremos verla realmente. Pero, luego, en lo que queda, encontramos otra piel que hay que eliminar, y así sucesivamente. Con este método, ¿llegamos con seguridad a la mente «real», o simplemente a la piel que no encierra nada? En tal caso toda mente es mecánica. (De todas formas, hemos explicado ya que no es una máquina de estado discreto).

Los últimos párrafos no pretenden ser argumentos convincentes, sino más bien deben tomarse como «una letanía destinada a inculcar una creencia».

El único apoyo realmente satisfactorio que puede darse a la opinión manifestada al principio del apartado 6 es el que consiste en esperar a finales de siglo y luego efectuar el experimento señalado. ¿Pero qué podemos decir entretanto? ¿Qué pasos hemos de dar ahora para que dé buen resultado el experimento?

Como he dicho, el problema fundamental estriba en programar. También serán

imprescindibles progresos de ingeniería, pero creo que estarán a la altura de las necesidades. Las estimaciones de la capacidad de almacenamiento del cerebro oscilan entre 10^{10} y 10^{15} dígitos binarios. Personalmente me inclino por el valor más bajo y creo que sólo una pequeña parte se utiliza para los tipos más elevados de pensamiento. La mayor parte de esta capacidad se emplea seguramente en la retención de impresiones visuales. Me sorprendería que se necesitara más de 10^9 para jugar bien al juego de imitación, en cualquier caso contra un hombre ciego. (Nota: la capacidad de la *Encyclopaedia Britannica*, décimoprimera edición, es de 2×10^9). Una capacidad de almacenamiento de 10^7 sería una posibilidad bastante real, aun con las técnicas actuales. Probablemente no será preciso aumentar la velocidad de operación de las máquinas. Partes de las máquinas modernas, que podríamos calificar de auténticas células nerviosas, trabajan mil veces más rápido que éstas. Con esto se conseguiría un «margen de seguridad» para compensar pérdidas de velocidad producidas por diversos motivos. El problema estriba, en último extremo, en saber cómo programar estas máquinas para jugar al juego. Con mi actual ritmo de trabajo produzco unos mil dígitos de programa diarios; en consecuencia, unas sesenta personas, trabajando asiduamente durante cincuenta años, podrían llevar a cabo esta tarea si no traspapelaran nada. Parece deseable un método más expeditivo.

En el proceso de intentar la imitación de una mente humana adulta estamos obligados a pensar muy en serio sobre el proceso por el que se ha llegado al estado en que se halla. Se observarán tres factores:

1. El estado inicial de la mente al nacer.
2. La educación que ha tenido.
3. Otras experiencias, aparte de la educación, a que haya estado sometida.

En lugar de intentar la elaboración de un programa que imite la mente adulta, ¿por qué no establecer uno que simule la mente infantil? Si luego la sometemos a un curso adecuado de formación, podría obtenerse un cerebro adulto. Podemos decir que el cerebro infantil es como el cuaderno recién comprado en una papelería: poco mecanismo y muchas hojas en blanco. (Mecanismo y escritura son casi sinónimos desde nuestro punto de vista). Nuestra esperanza se funda en que hay tan poco mecanismo en el cerebro infantil que debe resultar fácil programar algo similar. Podemos suponer que la cantidad de trabajo formativo, en una primera aproximación, sea muy parecida a la aplicable en el caso de un niño.

De este modo, el problema queda dividido en dos partes: el programa infantil y el proceso formativo. Ambos estrechamente vinculados. No puede esperarse construir una buena máquina infantil al primer intento; hay que experimentar enseñando a la máquina, y comprobar si aprende bien. Luego puede probarse otra vez y ver si es mejor o peor. Evidentemente existe una clara relación por analogía entre este proceso

y el de la evolución:

Estructura de la máquina infantil = Material hereditario
Cambios de la máquina infantil = Mutaciones
Selección natural = Juicio del experimentador

Sin embargo, es de esperar que este proceso sea más expeditivo que el de la evolución. La supervivencia del más apto es un método lento para valorar las ventajas. El experimentador, aplicando su inteligencia, debe ser capaz de acelerarlo. De igual importancia es el hecho de que no está limitado por mutaciones aleatorias. Si el experimentador descubre la causa de determinada debilidad, puede probablemente decidir el tipo de mutación que la mejore.

A la máquina no se le podrá aplicar exactamente el mismo proceso de aprendizaje que a un niño. Ya que, por ejemplo, no tendrá piernas y no se le podrá ordenar que vaya a por un cubo de carbón. Seguramente tampoco tendrá ojos. Y por mucho que se compensen estas deficiencias con una buena ingeniería, no se podrá enviar a la criatura a la escuela porque sería motivo de burla de sus compañeros. Habrá que darle clases particulares, sin preocuparnos por las piernas, los ojos, etc. El caso de Helen Keller demuestra que es posible la labor educativa a condición de que se establezca una comunicación bilateral entre maestro y alumno por el medio que sea.

Normalmente asociamos castigos y recompensas al proceso educativo. Algunas máquinas infantiles simples pueden construirse o programarse ateniéndose a ese principio. Hay que construir la máquina de tal modo que los acontecimientos que preceden brevemente a la aparición de la señal de castigo cuenten con mínimas posibilidades de repetición, y que, por el contrario, la señal de recompensa incremente la posibilidad de repetición de secuencias que la motivan. Estas especificaciones no presuponen tipo de sentimiento alguno por parte de la máquina. He realizado algunos experimentos con este tipo de máquina infantil y he logrado enseñarle varias cosas, pero utilicé un método de aprendizaje excesivamente heterodoxo para que el experimento pueda considerarse un éxito.

El empleo de castigos y recompensas puede a lo sumo formar parte del proceso de aprendizaje. En términos generales, si el enseñante no dispone de otros medios de comunicación con el alumno, la cantidad de información que éste recibe nunca excede el número de recompensas y castigos. Cuando un niño ha aprendido finalmente a repetir «Casabianca», se sentirá probablemente muy afligido si la única manera de dilucidar el texto es la técnica de las «Veinte preguntas» y cada «NO» supone una bofetada. Por lo tanto, es necesario disponer de otros canales de comunicación «no emocionales». Si los hay, se puede enseñar a una máquina por el método de premios y castigos a obedecer órdenes dadas en una lengua determinada, es decir un lenguaje simbólico. Estas órdenes se transmiten por canales «no

emocionales», y el empleo de dicho lenguaje disminuye notablemente la cantidad de castigos y premios.

Puede existir diversidad de opiniones en cuanto a la complejidad adecuada de la máquina infantil. Puede intentarse una construcción lo más simple posible, coherente con los principios generales. O puede dotársela de un sistema completo integrado de inferencia lógica, en cuyo caso el almacenamiento estará fundamentalmente ocupado por definiciones y proposiciones. Estas proposiciones serían de diversa índole: hechos bien establecidos, conjeturas, teoremas matemáticamente demostrables, afirmaciones hechas por una autoridad, expresiones con forma lógica de proposición pero de valor no creíble. Algunas proposiciones serían «imperativas». La máquina estaría construida de forma que, en cuanto una imperativa se clasificara como «bien establecida», se produjera automáticamente la acción apropiada. Como ejemplo, supongamos que el maestro dice a la máquina: «Ahora haz los deberes». Esto podría dar lugar a que «El maestro dice “Ahora haz los deberes”» quedara incluido en los hechos bien establecidos. Otra posibilidad sería: «Todo lo que dice el maestro es cierto». Ambas posibilidades combinadas podrían dar por resultado que la imperativa «Ahora haz los deberes» quedara incluida entre los hechos bien establecidos, lo cual, con arreglo a la construcción de la máquina, significaría que se inician realmente los deberes, pero el efecto es muy poco satisfactorio. El proceso de inferencia que utilice la máquina tiene que satisfacer al lógico más riguroso. Por ejemplo, no habrá jerarquía de tipos, lo que no significa que no se produzcan falacias de tipos, semejantes al riesgo de caer por un precipicio no señalizado. Unos imperativos adecuados (expresados *dentro* de los sistemas, pero que no formen parte de las reglas *del* sistema), tales como «No emplees una clase si no es una subclase de las mencionadas por el maestro», ejercerían la misma función que un letrero que indicara: «No acercarse al borde».

Las imperativas a las que obedece una máquina sin miembros son necesariamente de índole intelectual, como en el ejemplo citado (hacer los deberes). Entre dichas imperativas son importantes las que rigen el orden en que hay que aplicar las reglas del sistema lógico correspondiente, ya que, en cada fase de la utilización de un sistema lógico, hay una amplia alternativa de pasos que pueden seguirse para no transgredir las reglas de ese sistema lógico. Estas opciones marcan la diferencia entre un razonador brillante y otro torpe, pero no la diferencia entre uno serio y otro tramposo. Las proposiciones que conducen a las imperativas de esta clase pueden ser: «Cuando se mencione a Sócrates, utiliza el silogismo en Bárbara», o «Si se ha demostrado que un método es más rápido que otro, no uses el método lento». Algunas pueden «basarse en una autoridad», pero otras puede producirlas la propia máquina por inducción científica, por ejemplo.

La idea de una máquina que aprende puede parecer paradójica a algunos lectores.

¿Cómo pueden cambiarse las reglas de operación de la máquina? Estas deben especificar punto por punto cómo debe reaccionar la máquina independientemente de su historia y al margen de los cambios que experimente. Por lo tanto, las reglas son bastante invariables con respecto al tiempo. Y es bien cierto. La explicación de la paradoja consiste en que las reglas que cambian en el proceso de aprendizaje son de un tipo menos pretencioso y sólo tienen validez efímera. El lector puede establecer un paralelismo con la Constitución de los Estados Unidos.

Una característica importante de la máquina que aprende es la de que el profesor ignora muchas veces la mayoría de los procesos internos, aunque hasta cierto punto sea capaz de predecir el comportamiento de su alumno. Esto es tanto más aplicable a la formación ulterior de una máquina que tenga por origen una máquina infantil con un diseño (o programa) perfectamente experimentado. Situación muy distinta al procedimiento normal de emplear una máquina para hacer cálculos, ya que el objeto, en este caso, consiste en disponer de una imagen mental clara del estado de la máquina en cada momento de la computación. Este propósito sólo es alcanzable con una imposición. La opinión de que «la máquina sólo hace lo que queramos que haga parece extraña a la vista de lo expuesto. La mayoría de los programas que podemos introducir en la máquina la hará hacer algo que no entendemos o que consideramos como comportamiento totalmente aleatorio. El comportamiento inteligente consiste probablemente en una desviación del comportamiento absolutamente disciplinado que implica la computación, aunque relativamente leve y sin que provoque un comportamiento aleatorio o *loops* repetitivos inútiles. Otro importante resultado de la preparación de una máquina para que intervenga en el juego de imitación, merced a un proceso de enseñanza y aprendizaje, radica en que la «falibilidad humana» suele quedar descartada de una forma bastante natural, sin necesidad de «entrenamiento» especial. Los procesos que se aprenden no procuran una certeza absoluta de resultados; si así fuera, nunca fallaría su aprendizaje.

Quizá convenga introducir un elemento aleatorio en la máquina que aprende. Un elemento aleatorio resulta bastante útil en la búsqueda de la solución de un problema. Supongamos, por ejemplo, que deseamos hallar un número entre 50 y 200 que sea igual al cuadrado de la suma de sus cifras; empecemos por el 51 y sigamos con el 52 hasta encontrar la combinación justa. Otra alternativa sería elegir números al azar hasta hallar uno que nos sirva. Este método presenta la ventaja de que nos ahorra la necesidad de mantener el registro de los valores que se han probado, y el inconveniente de que se corre el riesgo de probar dos veces el mismo número, pero esto no es tan importante si hay varias soluciones. El método sistemático presenta el inconveniente de que puede haber una serie enorme sin solución en la región que hay que investigar en primer lugar. El proceso de aprendizaje puede considerarse como la búsqueda de una forma de comportamiento que satisfaga al profesor (o cualquier otro

requisito). Como probablemente existe un gran número de soluciones satisfactorias, el método aleatorio parece mejor que el sistemático. Se advertirá que es el que interviene en el proceso análogo de la evolución, y que en ella no es posible el método sistemático. ¿Cómo sería posible conservar el registro de las distintas combinaciones genéticas ensayadas para evitar probarlas de nuevo?

Esperemos que las máquinas lleguen a competir con el hombre en todos los campos puramente intelectuales. ¿Pero cuáles son los mejores para empezar? También es una ardua decisión. Muchos piensan que lo mejor es una actividad de naturaleza tan abstracta como jugar al ajedrez. También puede sostenerse que lo óptimo sería dotar a la máquina de los mejores órganos sensoriales posibles y luego enseñarla a entender y a hablar inglés. Es un proceso que podría hacerse con arreglo al aprendizaje normal de un niño: se señalan los objetos, se los nombra, etc. Vuelvo a insistir en que ignoro la respuesta adecuada; creo que hay que experimentar los dos enfoques.

Sólo podemos prever el futuro inmediato, pero de lo que no cabe duda es de que hay mucho por hacer.

INGLÉS

1. The Imitation Game

I propose to consider the question, «Can machines think?». This should begin with definitions of the meaning of the terms «machine» and «think». The definitions might be framed so as to reflect so far as possible the normal use of the words, but this attitude is dangerous. If the meaning of the words «machine» and «think» are to be found by examining how they are commonly used it is difficult to escape the conclusion that the meaning and the answer to the question, «Can machines think?» is to be sought in a statistical survey such as a Gallup poll. But this is absurd. Instead of attempting such a definition I shall replace the question by another, which is closely related to it and is expressed in relatively unambiguous words.

The new form of the problem can be described in terms of a game which we call the «imitation game». It is played with three people, a man (A), a woman (B), and an interrogator (C) who may be of either sex. The interrogator stays in a room apart front the other two. The object of the game for the interrogator is to determine which of the other two is the man and which is the woman. He knows them by labels X and Y, and at the end of the game he says either «X is A and Y is B» or «X is B and Y is A». The interrogator is allowed to put questions to A and B thus:

C: Will X please tell me the length of his or her hair?

Now suppose X is actually A, then A must answer. It is A’s object in the game to try and cause C to make the wrong identification. His answer might therefore be: «My hair is shingled, and the longest strands are about nine inches long».

In order that tones of voice may not help the interrogator the answers should be written, or better still, typewritten. The ideal arrangement is to have a teleprinter communicating between the two rooms. Alternatively the question and answers can be repeated by an intermediary. The object of the game for the third player (B) is to help the interrogator. The best strategy for her is probably to give truthful answers. She can add such things as «I am the woman, don’t listen to him!» to her answers, but it will avail nothing as the man can make similar remarks.

We now ask the question, «What will happen when a machine takes the part of A in this game?». Will the interrogator decide wrongly as often when the game is played like this as he does when the game is played between a man and a woman? These questions replace our original, «Can machines think?»

2. Critique of the New Problem

As well as asking, «What is the answer to this new form of the question», one may ask, «Is this new question a worthy one to investigate?». This latter question we investigate without further ado, thereby cutting short an infinite regress.

The new problem has the advantage of drawing a fairly sharp line between the physical and the intellectual capacities of a man. No engineer or chemist claims to be able to produce a material which is indistinguishable from the human skin. It is possible that at some time this might be done, but even supposing this invention available we should feel there was little point in trying to make a «thinking machine» more human by dressing it up in such artificial flesh. The form in which we have set the problem reflects this fact in the condition which prevents the interrogator from seeing or touching the other competitors, or hearing their voices. Some other advantages of the proposed criterion may be shown up by specimen questions and answers. Thus:

Q: Please write me a sonnet on the subject of the Forth Bridge.
A: Count me out on this one. I never could write poetry.
Q: Add 34957 to 70764.
A: (Pause about 30 seconds and then give as answer) 105621.
Q: Do you play chess?
A: Yes.
Q: I have K at my K1, and no other pieces. You have only K at K6 and R at R1. It is your move. What do you play?
A: (After a pause of 15 seconds) R-R8 mate.

The question and answer method seems to be suitable for introducing almost any one of the fields of human endeavour that we wish to include. We do not wish to penalise the machine for its inability to shine in beauty competitions, nor to penalise a man for losing in a race against an aeroplane. The conditions of our game make these disabilities irrelevant. The «witnesses» can brag, if they consider it advisable, as much as they please about their charms, strength or heroism, but the interrogator cannot demand practical demonstrations.

The game may perhaps be criticised on the ground that the odds are weighted too heavily against the machine. If the man were to try and pretend to be the machine he would clearly make a very poor showing. He would be given away at once by slowness and inaccuracy in arithmetic. May not machines carry out something which ought to be described as thinking but which is very different from what a man does? This objection is a very strong one, but at least we can say that if, nevertheless, a machine can be constructed to play the imitation game satisfactorily, we need not be

troubled by this objection.

It might be urged that when playing the «imitation game» the best strategy for the machine may possibly be something other than imitation of the behaviour of a man. This may be, but I think it is unlikely that there is any great effect of this kind. In any case there is no intention to investigate here the theory of the game, and it will be assumed that the best strategy is to try to provide answers that would naturally be given by a man.

3. The Machines Concerned in the Game

The question which we put in 1 will not be quite definite until we have specified what we mean by the word «machine». It is natural that we should wish to permit every kind of engineering technique to be used in our machines. We also wish to allow the possibility than an engineer or team of engineers may construct a machine which works, but whose manner of operation cannot be satisfactorily described by its constructors because they have applied a method which is largely experimental. Finally, we wish to exclude from the machines men born in the usual manner. It is difficult to frame the definitions so as to satisfy these three conditions. One might for instance insist that the team of engineers should be all of one sex, but this would not really be satisfactory, for it is probably possible to rear a complete individual from a single cell of the skin (say) of a man. To do so would be a feat of biological technique deserving of the very highest praise, but we would not be inclined to regard it as a case of «constructing a thinking machine». This prompts us to abandon the requirement that every kind of technique should be permitted. We are the more ready to do so in view of the fact that the present interest in «thinking machines» has been aroused by a particular kind of machine, usually called an «electronic computer» or «digital computer». Following this suggestion we only permit digital computers to take part in our game.

This restriction appears at first sight to be a very drastic one. I shall attempt to show that it is not so in reality. To do this necessitates a short account of the nature and properties of these computers. It may also be said that this identification of machines with digital computers, like our criterion for «thinking», will only be unsatisfactory if (contrary to my belief), it turns out that digital computers are unable to give a good showing in the game.

There are already a number of digital computers in working order, and it may be asked, «Why not try the experiment straight away? It would be easy to satisfy the conditions of the game. A number of interrogators could be used, and statistics compiled to show how often the right identification was given». The short answer is that we are not asking whether all digital computers would do well in the game nor whether the computers at present available would do well, but whether there are imaginable computers which would do well. But this is only the short answer. We shall see this question in a different light later.

4. Digital Computers

The idea behind digital computers may be explained by saying that these machines are intended to carry out any operations which could be done by a human computer. The human computer is supposed to be following fixed rules; he has no authority to deviate from them in any detail. We may suppose that these rules are supplied in a book, which is altered whenever he is put on to a new job. He has also an unlimited supply of paper on which he does his calculations. He may also do his multiplications and additions on a «desk machine», but this is not important.

If we use the above explanation as a definition we shall be in danger of circularity of argument. We avoid this by giving an outline. of the means by which the desired effect is achieved. A digital computer can usually be regarded as consisting of three parts:

1. (i) Store.
2. (ii) Executive unit.
3. (iii) Control.

The store is a store of information, and corresponds to the human computer's paper, whether this is the paper on which he does his calculations or that on which his book of rules is printed. In so far as the human computer does calculations in his bead a part of the store will correspond to his memory.

The executive unit is the part which carries out the various individual operations involved in a calculation. What these individual operations are will vary from machine to machine. Usually fairly lengthy operations can be done such as «Multiply 3540675445 by 7076345687» but in some machines only very simple ones such as «Write down 0» are possible.

We have mentioned that the «book of rules» supplied to the computer is replaced in the machine by a part of the store. It is then called the «table of instructions». It is the duty of the control to see that these instructions are obeyed correctly and in the right order. The control is so constructed that this necessarily happens.

The information in the store is usually broken up into packets of moderately small size. In one machine, for instance, a packet might consist of ten decimal digits. Numbers are assigned to the parts of the store in which the various packets of information are stored, in some systematic manner. A typical instruction might say: «Add the number stored in position 6809 to that in 4302 and put the result back into the latter storage position». Needless to say it would not occur in the machine expressed in English. It would more likely be coded in a form such as 6809430217. Here 17 says which of various possible operations is to be performed on the two numbers. In this case the operation is that described above, viz., «Add the

number…». It will be noticed that the instruction takes up 10 digits and so forms one packet of information, very conveniently. The control will normally take the instructions to be obeyed in the order of the positions in which they are stored, but occasionally an instruction such as: «Now obey the instruction stored in position 5606, and continue from there» may be encountered, or again: «If position 4505 contains 0 obey next the instruction stored in 6707, otherwise continue straight on».

Instructions of these latter types are very important because they make it possible for a sequence of operations to be replaced over and over again until some condition is fulfilled, but in doing so to obey, not fresh instructions on each repetition, but the same ones over and over again. To take a domestic analogy. Suppose Mother wants Tommy to call at the cobbler's every morning on his way to school to see if her shoes are done, she can ask him afresh every morning. Alternatively she can stick up a notice once and for all in the hall which he will see when he leaves for school and which tells him to call for the shoes, and also to destroy the notice when he comes back if he has the shoes with him.

The reader must accept it as a fact that digital computers can be constructed, and indeed have been constructed, according to the principles we have described, and that they can in fact mimic the actions of a human computer very closely.

The book of rules which we have described our human computer as using is of course a convenient fiction. Actual human computers really remember what they have got to do. If one wants to make a machine mimic the behaviour of the human computer in some complex operation one has to ask him how it is done, and then translate the answer into the form of an instruction table. Constructing instruction tables is usually described as «programming». To «programme a machine to carry out the operation A» means to put the appropriate instruction table into the machine so that it will do A.

An interesting variant on the idea of a digital computer is a «digital computer with a random element». These have instructions involving the throwing of a die or some equivalent electronic process; one such instruction might for instance be, «Throw the die and put the-resulting number into store 1000». Sometimes such a machine is described as having free will (though I would not use this phrase myself), It is not normally possible to determine from observing a machine whether it has a random element, for a similar effect can be produced by such devices as making the choices depend on the digits of the decimal for.

Most actual digital computers have only a finite store. There is no theoretical difficulty in the idea of a computer with an unlimited store. Of course only a finite part can have been used at any one time. Likewise only a finite amount can have been constructed, but we can imagine more and more being added as required. Such computers have special theoretical interest and will be called infinitive capacity

computers.

The idea of a digital computer is an old one. Charles Babbage, Lucasian Professor of Mathematics at Cambridge from 1828 to 1839, planned such a machine, called the Analytical Engine, but it was never completed. Although Babbage had all the essential ideas, his machine was not at that time such a very attractive prospect. The speed which would have been available would be definitely faster than a human computer but something like I 00 times slower than the Manchester machine, itself one of the slower of the modern machines, The storage was to be purely mechanical, using wheels and cards.

The fact that Babbage’s Analytical Engine was to be entirely mechanical will help us to rid ourselves of a superstition. Importance is often attached to the fact that modern digital computers are electrical, and that the nervous system also is electrical. Since Babbage’s machine was not electrical, and since all digital computers are in a sense equivalent, we see that this use of electricity cannot be of theoretical importance. Of course electricity usually comes in where fast signalling is concerned, so that it is not surprising that we find it in both these connections. In the nervous system chemical phenomena are at least as important as electrical. In certain computers the storage system is mainly acoustic. The feature of using electricity is thus seen to be only a very superficial similarity. If we wish to find such similarities we should took rather for mathematical analogies of function.

5. Universality of Digital Computers

The digital computers considered in the last section may be classified amongst the «discrete-state machines». These are the machines which move by sudden jumps or clicks from one quite definite state to another. These states are sufficiently different for the possibility of confusion between them to be ignored. Strictly speaking there, are no such machines. Everything really moves continuously. But there are many kinds of machine which can profitably be thought of as being discrete-state machines. For instance in considering the switches for a lighting system it is a convenient fiction that each switch must be definitely on or definitely off. There must be intermediate positions, but for most purposes we can forget about them. As an example of a discrete-state machine we might consider a wheel which clicks round through 120 once a second, but may be stopped by a]ever which can be operated from outside; in addition a lamp is to light in one of the positions of the wheel. This machine could be described abstractly as follows. The internal state of the machine (which is described by the position of the wheel) may be q_1, q_2 or q_3. There is an input signal i_0 or i_1 (position of lever). The internal state at any moment is determined by the last state and input signal according to the table:

	Last State		
Input	q_1	q_2	q_3
i_1	q_2	q_3	q_1
i_0	q_1	q_2	q_3

The output signals, the only externally visible indication of the internal state (the light) are described by the table:

State	q_1	q_2	q_3
Output	o_0	o_0	o_1

This example is typical of discrete-state machines. They can be described by such tables provided they have only a finite number of possible states.

It will seem that given the initial state of the machine and the input signals it is always possible to predict all future states, This is reminiscent of Laplace’s view that from the complete state of the universe at one moment of time, as described by the positions and velocities of all particles, it should be possible to predict all future states. The prediction which we are considering is, however, rather nearer to practicability than that considered by Laplace. The system of the «universe as a whole» is such that quite small errors in the initial conditions can have an

overwhelming effect at a later time. The displacement of a single electron by a billionth of a centimetre at one moment might make the difference between a man being killed by an avalanche a year later, or escaping. It is an essential property of the mechanical systems which we have called «discrete-state machines» that this phenomenon does not occur. Even when we consider the actual physical machines instead of the idealised machines, reasonably accurate knowledge of the state at one moment yields reasonably accurate knowledge any number of steps later.

As we have mentioned, digital computers fall within the class of discrete-state machines. But the number of states of which such a machine is capable is usually enormously large. For instance, the number for the machine now working at Manchester is about $2^{165,000}$, i.e., about $10^{50,000}$. Compare this with our example of the clicking wheel described above, which had three states. It is not difficult to see why the number of states should be so immense. The computer includes a store corresponding to the paper used by a human computer. It must be possible to write into the store any one of the combinations of symbols which might have been written on the paper. For simplicity suppose that only digits from 0 to 9 are used as symbols. Variations in handwriting are ignored. Suppose the computer is allowed 100 sheets of paper each containing 50 lines each with room for 30 digits. Then the number of states is $10^{100x50x30}$ i.e., $10^{150,000}$. This is about the number of states of three Manchester machines put together. The logarithm to the base two of the number of states is usually called the «storage capacity» of the machine. Thus the Manchester machine has a storage capacity of about 165,000 and the wheel machine of our example about 1.6. If two machines are put together their capacities must be added to obtain the capacity of the resultant machine. This leads to the possibility of statements such as «The Manchester machine contains 64 magnetic tracks each with a capacity of 2560, eight electronic tubes with a capacity of 1280. Miscellaneous storage amounts to about 300 making a total of 174,380».

Given the table corresponding to a discrete-state machine it is possible to predict what it will do. There is no reason why this calculation should not be carried out by means of a digital computer. Provided it could be carried out sufficiently quickly the digital computer could mimic the behavior of any discrete-state machine. The imitation game could then be played with the machine in question (as B) and the mimicking digital computer (as A) and the interrogator would be unable to distinguish them. Of course the digital computer must have an adequate storage capacity as well as working sufficiently fast. Moreover, it must be programmed afresh for each new machine which it is desired to mimic.

This special property of digital computers, that they can mimic any discrete-state machine, is described by saying that they are *universal* machines. The existence of machines with this property has the important consequence that, considerations of

speed apart, it is unnecessary to design various new machines to do various computing processes. They can all be done with one digital computer, suitably programmed for each case. It 'ill be seen that as a consequence of this all digital computers are in a sense equivalent.

We may now consider again the point raised at the end of §3. It was suggested tentatively that the question, «Can machines think?» should be replaced by «Are there imaginable digital computers which would do well in the imitation game?». If we wish we can make this superficially more general and ask «Are there discrete-state machines which would do well?». But in view of the universality property we see that either of these questions is equivalent to this, «Let us fix our attention on one particular digital computer C. Is it true that by modifying this computer to have an adequate storage, suitably increasing its speed of action, and providing it with an appropriate programme, C can be made to play satisfactorily the part of A in the imitation game, the part of B being taken by a man?».

6. Contrary Views on the Main Question

We may now consider the ground to have been cleared and we are ready to proceed to the debate on our question, «Can machines think?» and the variant of it quoted at the end of the last section. We cannot altogether abandon the original form of the problem, for opinions will differ as to the appropriateness of the substitution and we must at least listen to what has to be said in this connexion.

It will simplify matters for the reader if I explain first my own beliefs in the matter. Consider first the more accurate form of the question. I believe that in about fifty years' time it will be possible, to programme computers, with a storage capacity of about 10^9, to make them play the imitation game so well that an average interrogator will not have more than 70 per cent chance of making the right identification after five minutes of questioning. The original question, «Can machines think?». I believe to be too meaningless to deserve discussion. Nevertheless I believe that at the end of the century the use of words and general educated opinion will have altered so much that one will be able to speak of machines thinking without expecting to be contradicted. I believe further that no useful purpose is served by concealing these beliefs. The popular view that scientists proceed inexorably from well-established fact to well-established fact, never being influenced by any improved conjecture, is quite mistaken. Provided it is made clear which are proved facts and which are conjectures, no harm can result. Conjectures are of great importance since they suggest useful lines of research.

I now proceed to consider opinions opposed to my own.

(1) The Theological Objection

Thinking is a function of man's immortal soul. God has given an immortal soul to every man and woman, but not to any other animal or to machines. Hence no animal or machine can think.

I am unable to accept any part of this, but will attempt to reply in theological terms. I should find the argument more convincing if animals were classed with men, for there is a greater difference, to my mind, between the typical animate and the inanimate than there is between man and the other animals. The arbitrary character of the orthodox view becomes clearer if we consider how it might appear to a member of some other religious community. How do Christians regard the Moslem view that women have no souls? But let us leave this point aside and return to the main argument. It appears to me that the argument quoted above implies a serious restriction of the omnipotence of the Almighty. It is admitted that there are certain

things that He cannot do such as making one equal to two, but should we not believe that He has freedom to confer a soul on an elephant if He sees fit? We might expect that He would only exercise this power in conjunction with a mutation which provided the elephant with an appropriately improved brain to minister to the needs of this sort[. An argument of exactly similar form may be made for the case of machines. It may seem different because it is more difficult to «swallow». But this really only means that we think it would be less likely that He would consider the circumstances suitable for conferring a soul. The circumstances in question are discussed in the rest of this paper. In attempting to construct such machines we should not be irreverently usurping His power of creating souls, any more than we are in the procreation of children: rather we are, in either case, instruments of His will providing mansions for the souls that He creates.

However, this is mere speculation. I am not very impressed with theological arguments whatever they may be used to support. Such arguments have often been found unsatisfactory in the past. In the time of Galileo it was argued that the texts, «And the sun stood still… and hasted not to go down about a whole day» (Joshua x. 13) and «He laid the foundations of the earth, that it should not move at any time» (Psalm cv. 5) were an adequate refutation of the Copernican theory. With our present knowledge such an argument appears futile. When that knowledge was not available it made a quite different impression.

(2) The «Heads in the Sand» Objection

The consequences of machines thinking would be too dreadful. Let us hope and believe that they cannot do so."

This argument is seldom expressed quite so openly as in the form above. But it affects most of us who think about it at all. We like to believe that Man is in some subtle way superior to the rest of creation. It is best if he can be shown to be necessarily superior, for then there is no danger of him losing his commanding position. The popularity of the theological argument is clearly connected with this feeling. It is likely to be quite strong in intellectual people, since they value the power of thinking more highly than others, and are more inclined to base their belief in the superiority of Man on this power.

I do not think that this argument is sufficiently substantial to require refutation. Consolation would be more appropriate: perhaps this should be sought in the transmigration of souls.

(3) The Mathematical Objection

There are a number of results of mathematical logic which can be used to show that there are limitations to the powers of discrete-state machines. The best known of these results is known as Godel's theorem (1931) and shows that in any sufficiently powerful logical system statements can be formulated which can neither be proved nor disproved within the system, unless possibly the system itself is inconsistent. There are other, in some respects similar, results due to Church (1936), Kleene (1935), Rosser, and Turing (1937). The latter result is the most convenient to consider, since it refers directly to machines, whereas the others can only be used in a comparatively indirect argument: for instance if Godel's theorem is to be used we need in addition to have some means of describing logical systems in terms of machines, and machines in terms of logical systems. The result in question refers to a type of machine which is essentially a digital computer with an infinite capacity. It states that there are certain things that such a machine cannot do. If it is rigged up to give answers to questions as in the imitation game, there will be some questions to which it will either give a wrong answer, or fail to give an answer at all however much time is allowed for a reply. There may, of course, be many such questions, and questions which cannot be answered by one machine may be satisfactorily answered by another. We are of course supposing for the present that the questions are of the kind to which an answer «Yes» or «No» is appropriate, rather than questions such as «What do you think of Picasso?». The questions that we know the machines must fail on are of this type, «Consider the machine specified as follows… Will this machine ever answer "Yes" to any question?». The dots are to be replaced by a description of some machine in a standard form, which could be something like that used in §5. When the machine described bears a certain comparatively simple relation to the machine which is under interrogation, it can be shown that the answer is either wrong or not forthcoming. This is the mathematical result: it is argued that it proves a disability of machines to which the human intellect is not subject.

The short answer to this argument is that although it is established that there are limitations to the Powers If any particular machine, it has only been stated, without any sort of proof, that no such limitations apply to the human intellect. But I do not think this view can be dismissed quite so lightly. Whenever one of these machines is asked the appropriate critical question, and gives a definite answer, we know that this answer must be wrong, and this gives us a certain feeling of superiority. Is this feeling illusory? It is no doubt quite genuine, but I do not think too much importance should be attached to it. We too often give wrong answers to questions ourselves to be justified in being very pleased at such evidence of fallibility on the part of the machines. Further, our superiority can only be felt on such an occasion in relation to the one machine over which we have scored our petty triumph. There would be no question of triumphing simultaneously over *all* machines. In short, then, there might

be men cleverer than any given machine, but then again there might be other machines cleverer again, and so on.

Those who hold to the mathematical argument would, I think, mostly he willing to accept the imitation game as a basis for discussion, Those who believe in the two previous objections would probably not be interested in any criteria.

(4) The Argument from Consciousness

This argument is very, well expressed in Professor Jefferson's Lister Oration for 1949, from which I quote. «Not until a machine can write a sonnet or compose a concerto because of thoughts and emotions felt, and not by the chance fall of symbols, could we agree that machine equals brain-that is, not only write it but know that it had written it. No mechanism could feel (and not merely artificially signal, an easy contrivance) pleasure at its successes, grief when its valves fuse, be warmed by flattery, be made miserable by its mistakes, be charmed by sex, be angry or depressed when it cannot get what it wants».

This argument appears to be a denial of the validity of our test. According to the most extreme form of this view the only way by which one could be sure that machine thinks is to be the machine and to feel oneself thinking. One could then describe these feelings to the world, but of course no one would be justified in taking any notice. Likewise according to this view the only way to know that a man thinks is to be that particular man. It is in fact the solipsist point of view. It may be the most logical view to hold but it makes communication of ideas difficult. A is liable to believe «A thinks but B does not» whilst B believes «B thinks but A does not». Instead of arguing continually over this point it is usual to have the polite convention that everyone thinks.

I am sure that Professor Jefferson does not wish to adopt the extreme and solipsist point of view. Probably he would be quite willing to accept the imitation game as a test. The game (with the player B omitted) is frequently used in practice under the name of *viva voce* to discover whether some one really understands something or has «learnt it parrot fashion». Let us listen in to a part of such a *viva voce*:

Interrogator: In the first line of your sonnet which reads «Shall I compare thee to a summer's day», would not «a spring day» do as well or better?

Witness: It wouldn't scan.

Interrogator: How about «a winter's day»? That would scan all right.

Witness: Yes, but nobody wants to be compared to a winter's day.

Interrogator: Would you say Mr. Pickwick reminded you of Christmas?

Witness: In a way.

Interrogator: Yet Christmas is a winter's day, and I do not think Mr. Pickwick would mind the comparison.

Witness: I don't think you're serious. By a winter's day one means a typical winter's day, rather than a special one like Christmas.

And so on, What would Professor Jefferson say if the sonnet-writing machine was able to answer like this in the viva voce? I do not know whether he would regard the machine as «merely artificially signalling» these answers, but if the answers were as satisfactory and sustained as in the above passage I do not think he would describe it as «an easy contrivance». This phrase is, I think, intended to cover such devices as the inclusion in the machine of a record of someone reading a sonnet, with appropriate switching to turn it on from time to time.

In short then, I think that most of those who support the argument from consciousness could be persuaded to abandon it rather than be forced into the solipsist position. They will then probably be willing to accept our test.

I do not wish to give the impression that I think there is no mystery about consciousness. There is, for instance, something of a paradox connected with any attempt to localise it. But I do not think these mysteries necessarily need to be solved before we can answer the question with which we are concerned in this paper.

(5) Arguments from Various Disabilities

These arguments take the form, «I grant you that you can make machines do all the things you have mentioned but you will never be able to make one to do X». Numerous features X are suggested in this connexion I offer a selection:

Be kind, resourceful, beautiful, friendly, have initiative, have a sense of humour, tell right from wrong, make mistakes, fall in love, enjoy strawberries and cream, make some one fall in love with it, learn from experience, use words properly, be the subject of its own thought, have as much diversity of behaviour as a man, do something really new.

No support is usually offered for these statements. I believe they are mostly founded on the principle of scientific induction. A man has seen thousands of machines in his lifetime. From what he sees of them he draws a number of general conclusions. They are ugly, each is designed for a very limited purpose, when required for a minutely different purpose they are useless, the variety of behaviour of any one of them is very small, etc., etc. Naturally he concludes that these are necessary properties of machines in general. Many of these limitations are associated with the very small storage capacity of most machines. (I am assuming that the idea of storage capacity is extended in some way to cover machines other than discrete-state machines. The exact definition does not matter as no mathematical accuracy is

claimed in the present discussion). A few years ago, when very little had been heard of digital computers, it was possible to elicit much incredulity concerning them, if one mentioned their properties without describing their construction. That was presumably due to a similar application of the principle of scientific induction. These applications of the principle are of course largely unconscious. When a burnt child fears the fire and shows that he fears it by avoiding it, f should say that he was applying scientific induction. (I could of course also describe his behaviour in many other ways). The works and customs of mankind do not seem to be very suitable material to which to apply scientific induction. A very large part of space-time must be investigated, if reliable results are to be obtained. Otherwise we may (as most English Children do) decide that everybody speaks English, and that it is silly to learn French.

There are, however, special remarks to be made about many of the disabilities that have been mentioned. The inability to enjoy strawberries and cream may have struck the reader as frivolous. Possibly a machine might be made to enjoy this delicious dish, but any attempt to make one do so would be idiotic. What is important about this disability is that it contributes to some of the other disabilities, e.g., to the difficulty of the same kind of friendliness occurring between man and machine as between white man and white man, or between black man and black man.

The claim that «machines cannot make mistakes» seems a curious one. One is tempted to retort, «Are they any the worse for that?». But let us adopt a more sympathetic attitude, and try to see what is really meant. I think this criticism can be explained in terms of the imitation game. It is claimed that the interrogator could distinguish the machine from the man simply by setting them a number of problems in arithmetic. The machine would be unmasked because of its deadly accuracy. The reply to this is simple. The machine (programmed for playing the game) would not attempt to give the *right* answers to the arithmetic problems. It would deliberately introduce mistakes in a manner calculated to confuse the interrogator. A mechanical fault would probably show itself through an unsuitable decision as to what sort of a mistake to make in the arithmetic. Even this interpretation of the criticism is not sufficiently sympathetic. But we cannot afford the space to go into it much further. It seems to me that this criticism depends on a confusion between two kinds of mistake, We may call them «errors of functioning» and «errors of conclusion». Errors of functioning are due to some mechanical or electrical fault which causes the machine to behave otherwise than it was designed to do. In philosophical discussions one likes to ignore the possibility of such errors; one is therefore discussing «abstract machines». These abstract machines are mathematical fictions rather than physical objects. By definition they are incapable of errors of functioning. In this sense we can truly say that «machines can never make mistakes». Errors of conclusion can only

arise when some meaning is attached to the output signals from the machine. The machine might, for instance, type out mathematical equations, or sentences in English. When a false proposition is typed we say that the machine has committed an error of conclusion. There is clearly no reason at all for saying that a machine cannot make this kind of mistake. It might do nothing but type out repeatedly «O = I». To take a less perverse example, it might have some method for drawing conclusions by scientific induction. We must expect such a method to lead occasionally to erroneous results.

The claim that a machine cannot be the subject of its own thought can of course only be answered if it can be shown that the machine has some thought with *some* subject matter. Nevertheless, «the subject matter of a machine's operations» does seem to mean something, at least to the people who deal with it. If, for instance, the machine was trying to find a solution of the equation $x^2 - 40x - 11 = 0$ one would be tempted to describe this equation as part of the machine's subject matter at that moment. In this sort of sense a machine undoubtedly can be its own subject matter. It may be used to help in making up its own programmes, or to predict the effect of alterations in its own structure. By observing the results of its own behaviour it can modify its own programmes so as to achieve some purpose more effectively. These are possibilities of the near future, rather than Utopian dreams.

The criticism that a machine cannot have much diversity of behaviour is just a way of saying that it cannot have much storage capacity. Until fairly recently a storage capacity of even a thousand digits was very rare.

The criticisms that we are considering here are often disguised forms of the argument from consciousness, Usually if one maintains that a machine can do one of these things, and describes the kind of method that the machine *could* use, one will not make much of an impression. It is thought that tile method (whatever it may be, for it must be mechanical) is really rather base. Compare the parentheses in Jefferson's statement quoted on page 22.

(6) Lady Lovelace's Objection

Our most detailed information of Babbage's Analytical Engine comes from a memoir by Lady Lovelace (1842). In it she states, «The Analytical Engine has no pretensions to *originate* anything. It can do *whatever we know how to order it to perform*» (her italics). This statement is quoted by Hartree (1949) who adds: «This does not imply that it may not be possible to construct electronic equipment which will "think for itself" or in which, in biological terms, one could set up a conditioned reflex, which would serve as a basis for "learning". Whether this is possible in principle or not is a stimulating and exciting question, suggested by some of these

recent developments But it did not seem that the machines constructed or projected at the time had this property».

I am in thorough agreement with Hartree over this. It will be noticed that he does not assert that the machines in question had not got the property, but rather that the evidence available to Lady Lovelace did not encourage her to believe that they had it. It is quite possible that the machines in question had in a sense got this property. For suppose that some discrete-state machine has the property. The Analytical Engine was a universal digital computer, so that, if its storage capacity and speed were adequate, it could by suitable programming be made to mimic the machine in question. Probably this argument did not occur to the Countess or to Babbage. In any case there was no obligation on them to claim all that could be claimed.

This whole question will be considered again under the heading of learning machines.

A variant of Lady Lovelace's objection states that a machine can «never do anything really new». This may be parried for a moment with the saw, «There is nothing new under the sun». Who can be certain that «original work» that he has done was not simply the growth of the seed planted in him by teaching, or the effect of following well-known general principles. A better variant of the objection says that a machine can never «take us by surprise». This statement is a more direct challenge and can be met directly. Machines take me by surprise with great frequency. This is largely because I do not do sufficient calculation to decide what to expect them to do, or rather because, although I do a calculation, I do it in a hurried, slipshod fashion, taking risks. Perhaps I say to myself, «I suppose the Voltage here ought to he the same as there: anyway let's assume it is». Naturally I am often wrong, and the result is a surprise for me for by the time the experiment is done these assumptions have been forgotten. These admissions lay me open to lectures on the subject of my vicious ways, but do not throw any doubt on my credibility when I testify to the surprises I experience.

I do not expect this reply to silence my critic. He will probably say that h surprises are due to some creative mental act on my part, and reflect no credit on the machine. This leads us back to the argument from consciousness, and far from the idea of surprise. It is a line of argument we must consider closed, but it is perhaps worth remarking that the appreciation of something as surprising requires as much of a «creative mental act» whether the surprising event originates from a man, a book, a machine or anything else.

The view that machines cannot give rise to surprises is due, I believe, to a fallacy to which philosophers and mathematicians are particularly subject. This is the assumption that as soon as a fact is presented to a mind all consequences of that fact spring into the mind simultaneously with it. It is a very useful assumption under

many circumstances, but one too easily forgets that it is false. A natural consequence of doing so is that one then assumes that there is no virtue in the mere working out of consequences from data and general principles.

(7) Argument from Continuity in the Nervous System

The nervous system is certainly not a discrete-state machine. A small error in the information about the size of a nervous impulse impinging on a neuron, may make a large difference to the size of the outgoing impulse. It may be argued that, this being so, one cannot expect to be able to mimic the behaviour of the nervous system with a discrete-state system.

It is true that a discrete-state machine must be different from a continuous machine. But if we adhere to the conditions of the imitation game, the interrogator will not be able to take any advantage of this difference. The situation can be made clearer if we consider sonic other simpler continuous machine. A differential analyser will do very well. (A differential analyser is a certain kind of machine not of the discrete-state type used for some kinds of calculation). Some of these provide their answers in a typed form, and so are suitable for taking part in the game. It would not be possible for a digital computer to predict exactly what answers the differential analyser would give to a problem, but it would be quite capable of giving the right sort of answer. For instance, if asked to give the value of (actually about 3.1416) it would be reasonable to choose at random between the values 3.12, 3.13, 3.14, 3.15, 3.16 with the probabilities of 0.05, 0.15, 0.55, 0.19, 0.06 (say). Under these circumstances it would be very difficult for the interrogator to distinguish the differential analyser from the digital computer.

(8) The Argument from Informality of Behaviour

It is not possible to produce a set of rules purporting to describe what a man should do in every conceivable set of circumstances. One might for instance have a rule that one is to stop when one sees a red traffic light, and to go if one sees a green one, but what if by some fault both appear together? One may perhaps decide that it is safest to stop. But some further difficulty may well arise from this decision later. To attempt to provide rules of conduct to cover every eventuality, even those arising from traffic lights, appears to be impossible. With all this I agree.

From this it is argued that we cannot be machines. I shall try to reproduce the argument, but I fear I shall hardly do it justice. It seems to run something like this: «if each man had a definite set of rules of conduct by which he regulated his life he

would be no better than a machine. But there are no such rules, so men cannot be machines». The undistributed middle is glaring. I do not think the argument is ever put quite like this, but I believe this is the argument used nevertheless. There may however be a certain confusion between «rules of conduct» and «laws of behaviour» to cloud the issue. By «rules of conduct» I mean precepts such as «Stop if you see red lights», on which one can act, and of which one can be conscious. By «laws of behaviour» I mean laws of nature as applied to a man's body such as «if you pinch him he will squeak». If we substitute «laws of behaviour which regulate his life» for «laws of conduct by which he regulates his life» in the argument quoted the undistributed middle is no longer insuperable. For we believe that it is not only true that being regulated by laws of behaviour implies being some sort of machine (though not necessarily a discrete-state machine), but that conversely being such a machine implies being regulated by such laws. However, we cannot so easily convince ourselves of the absence of complete laws of behaviour as of complete rules of conduct. The only way we know of for finding such laws is scientific observation, and we certainly know of no circumstances under which we could say, «We have searched enough. There are no such laws».

We can demonstrate more forcibly that any such statement would be unjustified. For suppose we could be sure of finding such laws if they existed. Then given a discrete-state machine it should certainly be possible to discover by observation sufficient about it to predict its future behaviour, and this within a reasonable time, say a thousand years. But this does not seem to be the case. I have set up on the Manchester computer a small programme using only 1,000 units of storage, whereby the machine supplied with one sixteen-figure number replies with another within two seconds. I would defy anyone to learn from these replies sufficient about the programme to be able to predict any replies to untried values.

(9) The Argument from Extrasensory Perception

I assume that the reader is familiar with the idea of extrasensory perception, and the meaning of the four items of it, viz., telepathy, clairvoyance, precognition and psychokinesis. These disturbing phenomena seem to deny all our usual scientific ideas. How we should like to discredit them! Unfortunately the statistical evidence, at least for telepathy, is overwhelming. It is very difficult to rearrange one's ideas so as to fit these new facts in. Once one has accepted them it does not seem a very big step to believe in ghosts and bogies. The idea that our bodies move simply according to the known laws of physics, together with some others not yet discovered but somewhat similar, would be one of the first to go.

This argument is to my mind quite a strong one. One can say in reply that many

scientific theories seem to remain workable in practice, in spite of clashing with ESP; that in fact one can get along very nicely if one forgets about it. This is rather cold comfort, and one fears that thinking is just the kind of phenomenon where ESP may be especially relevant.

A more specific argument based on ESP might run as follows: «Let us play the imitation game, using as witnesses a man who is good as a telepathic receiver, and a digital computer. The interrogator can ask such questions as “What suit does the card in my right hand belong to?”. The man by telepathy or clairvoyance gives the right answer 130 times out of 400 cards. The machine can only guess at random, and perhaps gets 104 right, so the interrogator makes the right identification». There is an interesting possibility which opens here. Suppose the digital computer contains a random number generator. Then it will be natural to use this to decide what answer to give. But then the random number generator will be subject to the psychokinetic powers of the interrogator. Perhaps this psychokinesis might cause the machine to guess right more often than would be expected on a probability calculation, so that the interrogator might still be unable to make the right identification. On the other hand, he might be able to guess right without any questioning, by clairvoyance. With ESP anything may happen.

If telepathy is admitted it will be necessary to tighten our test up. The situation could be regarded as analogous to that which would occur if the interrogator were talking to himself and one of the competitors was listening with his ear to the wall. To put the competitors into a «telepathy-proof room» would satisfy all requirements.

7. Learning Machines

The reader will have anticipated that I have no very convincing arguments of a positive nature to support my views. If I had I should not have taken such pains to point out the fallacies in contrary views. Such evidence as I have I shall now give.

Let us return for a moment to Lady Lovelace's objection, which stated that the machine can only do what we tell it to do. One could say that a man can «inject» an idea into the machine, and that it will respond to a certain extent and then drop into quiescence, like a piano string struck by a hammer. Another simile would be an atomic pile of less than critical size: an injected idea is to correspond to a neutron entering the pile from without. Each such neutron will cause a certain disturbance which eventually dies away. If, however, the size of the pile is sufficiently increased, tire disturbance caused by such an incoming neutron will very likely go on and on increasing until the whole pile is destroyed. Is there a corresponding phenomenon for minds, and is there one for machines? There does seem to be one for the human mind. The majority of them seem to be «subcritical», i.e., to correspond in this analogy to piles of subcritical size. An idea presented to such a mind will on average give rise to less than one idea in reply. A smallish proportion are supercritical. An idea presented to such a mind that may give rise to a whole «theory» consisting of secondary, tertiary and more remote ideas. Animals minds seem to be very definitely subcritical. Adhering to this analogy we ask, «Can a machine be made to be supercritical?»

The «skin-of-an-onion» analogy is also helpful. In considering the functions of the mind or the brain we find certain operations which we can explain in purely mechanical terms. This we say does not correspond to the real mind: it is a sort of skin which we must strip off if we are to find the real mind. But then in what remains we find a further skin to be stripped off, and so on. Proceeding in this way do we ever come to the «real» mind, or do we eventually come to the skin which has nothing in it? In the latter case the whole mind is mechanical. (It would not be a discrete-state machine however. We have discussed this).

These last two paragraphs do not claim to be convincing arguments. They should rather be described as «recitations tending to produce belief».

The only really satisfactory support that can be given for the view expressed at the beginning of §6, will be that provided by waiting for the end of the century and then doing the experiment described. But what can we say in the meantime? What steps should be taken now if the experiment is to be successful?

As I have explained, the problem is mainly one of programming. Advances in engineering will have to be made too, but it seems unlikely that these will not be adequate for the requirements. Estimates of the storage capacity of the brain vary

from 10^{10} to 10^{15} binary digits. I incline to the lower values and believe that only a very small fraction is used for the higher types of thinking. Most of it is probably used for the retention of visual impressions, I should be surprised if more than 10^9 was required for satisfactory playing of the imitation game, at any rate against a blind man. (Note: The capacity of the *Encyclopaedia Britannica*, 11th edition, is 2 X 10^9). A storage capacity of 10^7, would be a very practicable possibility even by present techniques. It is probably not necessary to increase the speed of operations of the machines at all. Parts of modern machines which can be regarded as analogs of nerve cells work about a thousand times faster than the latter. This should provide a «margin of safety» which could cover losses of speed arising in many ways, Our problem then is to find out how to programme these machines to play the game. At my present rate of working I produce about a thousand digits of progratiirne a day, so that about sixty workers, working steadily through the fifty years might accomplish the job, if nothing went into the wastepaper basket. Some more expeditious method seems desirable.

In the process of trying to imitate an adult human mind we are bound to think a good deal about the process which has brought it to the state that it is in. We may notice three components:

(a) The initial state of the mind, say at birth,
(b) The education to which it has been subjected,
(c) Other experience, not to be described as education, to which it has been subjected.

Instead of trying to produce a programme to simulate the adult mind, why not rather try to produce one which simulates the child's? If this were then subjected to an appropriate course of education one would obtain the adult brain. Presumably the child brain is something like a notebook as one buys it from the stationer's. Rather little mechanism, and lots of blank sheets. (Mechanism and writing are from our point of view almost synonymous). Our hope is that there is so little mechanism in the child brain that something like it can be easily programmed. The amount of work in the education we can assume, as a first approximation, to be much the same as for the human child.

We have thus divided our problem into two parts. The child programme and the education process. These two remain very closely connected. We cannot expect to find a good child machine at the first attempt. One must experiment with teaching one such machine and see how well it learns. One can then try another and see if it is better or worse. There is an obvious connection between this process and evolution, by the identifications

Structure of the child machine = hereditary material
Changes of the child machine = mutation,
Natural selection = judgment of the experimenter

One may hope, however, that this process will be more expeditious than evolution. The survival of the fittest is a slow method for measuring advantages. The experimenter, by the exercise of intelligence, should he able to speed it up. Equally important is the fact that he is not restricted to random mutations. If he can trace a cause for some weakness he can probably think of the kind of mutation which will improve it.

It will not be possible to apply exactly the same teaching process to the machine as to a normal child. It will not, for instance, be provided with legs, so that it could not be asked to go out and fill the coal scuttle. Possibly it might not have eyes. But however well these deficiencies might be overcome by clever engineering, one could not send the creature to school without the other children making excessive fun of it. It must be given some tuition. We need not be too concerned about the legs, eyes, etc. The example of Miss Helen Keller shows that education can take place provided that communication in both directions between teacher and pupil can take place by some means or other.

We normally associate punishments and rewards with the teaching process. Some simple child machines can be constructed or programmed on this sort of principle. The machine has to be so constructed that events which shortly preceded the occurrence of a punishment signal are unlikely to be repeated, whereas a reward signal increased the probability of repetition of the events which led up to it. These definitions do not presuppose any feelings on the part of the machine, I have done some experiments with one such child machine, and succeeded in teaching it a few things, but the teaching method was too unorthodox for the experiment to be considered really successful.

The use of punishments and rewards can at best be a part of the teaching process. Roughly speaking, if the teacher has no other means of communicating to the pupil, the amount of information which can reach him does not exceed the total number of rewards and punishments applied. By the time a child has learnt to repeat «Casabianca» he would probably feel very sore indeed, if the text could only be discovered by a «Twenty Questions» technique, every «NO» taking the form of a blow. It is necessary therefore to have some other «unemotional» channels of communication. If these are available it is possible to teach a machine by punishments and rewards to obey orders given in some language, e.g., a symbolic language. These orders are to be transmitted through the «unemotional» channels. The use of this language will diminish greatly the number of punishments and

rewards required.

Opinions may vary as to the complexity which is suitable in the child machine. One might try to make it as simple as possible consistently with the general principles. Alternatively one might have a complete system of logical inference «built in». In the latter case the store would be largely occupied with definitions and propositions. The propositions would have various kinds of status, e.g., well-established facts, conjectures, mathematically proved theorems, statements given by an authority, expressions having the logical form of proposition but not belief-value. Certain propositions may be described as «imperatives». The machine should be so constructed that as soon as an imperative is classed as «well established» the appropriate action automatically takes place. To illustrate this, suppose the teacher says to the machine, «Do your homework now». This may cause «Teacher says "Do your homework now"» to be included amongst the well-established facts. Another such fact might be, «Everything that teacher says is true». Combining these may eventually lead to the imperative, «Do your homework now», being included amongst the well-established facts, and this, by the construction of the machine, will mean that the homework actually gets started, but the effect is very satisfactory. The processes of inference used by the machine need not be such as would satisfy the most exacting logicians. There might for instance be no hierarchy of types. But this need not mean that type fallacies will occur, any more than we are bound to fall over unfenced cliffs. Suitable imperatives (expressed *within* the systems, not forming part of the rules *of the* system) such as «Do not use a class unless it is a subclass of one which has been mentioned by teacher» can have a similar effect to «Do not go too near the edge».

The imperatives that can be obeyed by a machine that has no limbs are bound to be of a rather intellectual character, as in the example (doing homework) given above. important amongst such imperatives will be ones which regulate the order in which the rules of the logical system concerned are to be applied, For at each stage when one is using a logical system, there is a very large number of alternative steps, any of which one is permitted to apply, so far as obedience to the rules of the logical system is concerned. These choices make the difference between a brilliant and a footling reasoner, not the difference between a sound and a fallacious one. Propositions leading to imperatives of this kind might be «When Socrates is mentioned, use the syllogism in Barbara» or «If one method has been proved to be quicker than another, do not use the slower method». Some of these may be «given by authority», but others may be produced by the machine itself, e.g. by scientific induction.

The idea of a learning machine may appear paradoxical to some readers. How can the rules of operation of the machine change? They should describe completely how

the machine will react whatever its history might be, whatever changes it might undergo. The rules are thus quite time-invariant. This is quite true. The explanation of the paradox is that the rules which get changed in the learning process are of a rather less pretentious kind, claiming only an ephemeral validity. The reader may draw a parallel with the Constitution of the United States.

An important feature of a learning machine is that its teacher will often be very largely ignorant of quite what is going on inside, although he may still be able to some extent to predict his pupil's behavior. This should apply most strongly to the later education of a machine arising from a child machine of well-tried design (or programme). This is in clear contrast with normal procedure when using a machine to do computations one's object is then to have a clear mental picture of the state of the machine at each moment in the computation. This object can only be achieved with a struggle. The view that «the machine can only do what we know how to order it to do», appears strange in face of this. Most of the programmes which we can put into the machine will result in its doing something that we cannot make sense (if at all, or which we regard as completely random behaviour. Intelligent behaviour presumably consists in a departure from the completely disciplined behaviour involved in computation, but a rather slight one, which does not give rise to random behaviour, or to pointless repetitive *loops*. Another important result of preparing our machine for its part in the imitation game by a process of teaching and learning is that «human fallibility» is likely to be omitted in a rather natural way, i.e., without special «coaching». (The reader should reconcile this with the point of view on pages 23 and 24). Processes that are learnt do not produce a hundred per cent certainty of result; if they did they could not be unlearnt.

It is probably wise to include a random element in a learning machine. A random element is rather useful when we are searching for a solution of some problem. Suppose for instance we wanted to find a number between 50 and 200 which was equal to the square of the sum of its digits, we might start at 51 then try 52 and go on until we got a number that worked. Alternatively we might choose numbers at random until we got a good one. This method has the advantage that it is unnecessary to keep track of the values that have been tried, but the disadvantage that one may try the same one twice, but this is not very important if there are several solutions. The systematic method has the disadvantage that there may be an enormous block without any solutions in the region which has to be investigated first, Now the learning process may be regarded as a search for a form of behaviour which will satisfy the teacher (or some other criterion). Since there is probably a very large number of satisfactory solutions the random method seems to be better than the systematic. It should be noticed that it is used in the analogous process of evolution. But there the systematic method is not possible. How could one keep track of the different

genetical combinations that had been tried, so as to avoid trying them again?

We may hope that machines will eventually compete with men in all purely intellectual fields. But which are the best ones to start with? Even this is a difficult decision. Many people think that a very abstract activity, like the playing of chess, would be best. It can also be maintained that it is best to provide the machine with the best sense organs that money can buy, and then teach it to understand and speak English. This process could follow the normal teaching of a child. Things would be pointed out and named, etc. Again I do not know what the right answer is, but I think both approaches should be tried.

We can only see a short distance ahead, but we can see plenty there that needs to be done.

ALAN MATHISON TURING (1912 - 1954). Nació en Londres (Gran Bretaña), desde muy temprana edad Turing demostró su inteligencia. A los 3 años tenía una inusual capacidad para recordar palabras y a los 8 años se interesó por la química montando un laboratorio en su casa. Con 13 años ingresó en la escuela Sherborne, en la que ya demostraba su facilidad para las matemáticas, teniendo una gran capacidad para realizar cálculos mentalmente.

Obtuvo una beca para estudiar en la universidad de Cambridge, en donde se graduó de la licenciatura de matemáticas con honores en 1934. En abril de 1936, publicó el artículo «*On computable numbers, with an application to the Entscheidungsproblem*» en el que introduce el concepto de algoritmo y de máquina de Turing. Este artículo da respuesta (negativa) al problema de la decisión formulada por Hilbert en 1900, probando que existen problemas sin solución algorítmica y es uno de los cimientos más importantes de la teoría de la computación.

En septiembre de 1936, Turing ingresó en la universidad de Princeton (EE.UU.). Su artículo atrajo la atención de uno de los científicos más destacados de la época, John von Neumann, quien le ofreció una beca en el Instituto de Estudios Avanzados. Turing obtuvo su doctorado en matemáticas en 1938. Tras su graduación, von Neumann le ofreció una plaza como su asistente, pero Turing rechazó la oferta y volvió a Inglaterra, en donde vivió de una beca universitaria mientras estudiaba filosofía de las matemáticas entre 1938 y 1939.

En 1939, con el comienzo de la Segunda Guerra Mundial, Turing fue reclutado por el

ejército británico para descifrar los códigos emitidos por la máquina Enigma utilizada por los alemanes. En el deseo de obtener mejores máquinas descifradoras, se comenzó a construir la primera computadora electrónica, llamada Colossus, bajo la supervisión de Turing, se construyeron 10 unidades, y la primera empezó a operar en 1943. Por su trabajo en el Colossus, Turing recibió la Orden del Imperio Británico en 1946.

En 1944, Turing fue contratado por el Laboratorio Nacional de Física (NLP) para competir con el proyecto americano EDVAC, de von Neumann. Turing ejerció como Oficial Científico Principal a cargo del Automatic Computing Engine (ACE). Hacia 1947, Turing concibió la idea de las redes de cómputo y el concepto de subrutina y biblioteca de software. También describió las ideas básicas de lo que hoy se conoce como red neuronal. Abandonó la NLP en 1948.

Turing se adelantó al proyecto de construcción de un ordenador de acuerdo con la arquitectura de von Neumann. El Manchester Mark I, estuvo acabado en 1948 antes que el EDVAC. Turing diseñó para esta máquina un lenguaje de programación basado en el código empleado por los teletipos.

Otro de los campos de investigación de Turing fue la inteligencia artificial, se puede decir que esta disciplina nació a partir del artículo titulado «*Computing Machinery and Inteligence*» publicado por Turing en 1950. Es muy famosa la primera frase de este artículo: «Propongo considerar la siguiente cuestión: ¿Pueden pensar las máquinas?». Turing propuso un método llamado el test de Turing para determinar si las máquinas podrían tener la capacidad de pensar.

En 1951, es nombrado miembro de la Sociedad Real de Londres por sus contribuciones científicas. Y en su honor, la Association for Computing Machinery llama «Turing Award» a su premio más importante, el cual se otorga desde 1966 a los expertos que han realizado las mayores contribuciones al avance de la computación.