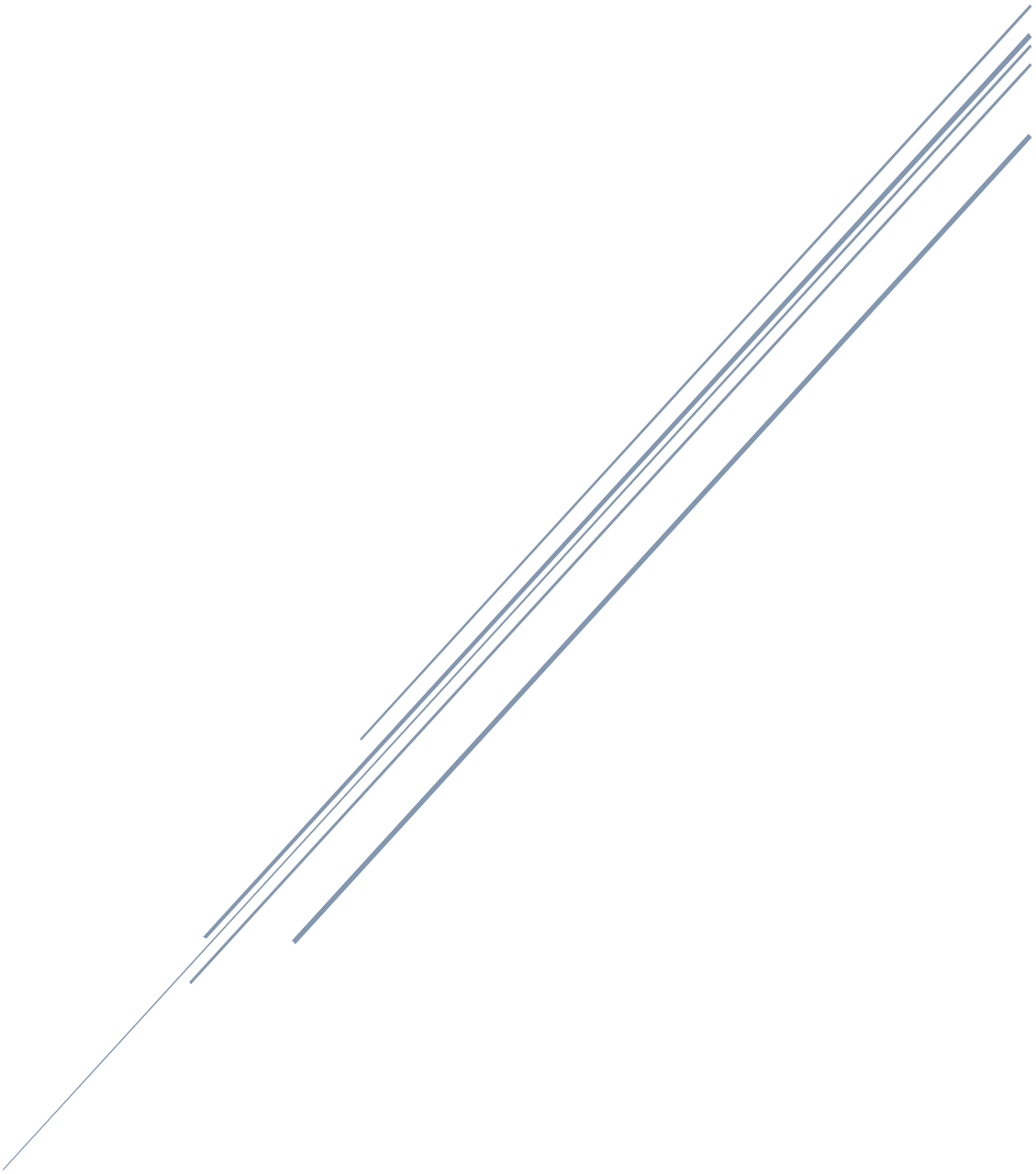
# ACTIVIDAD 2 – UNIDAD 1

## Unidad 1



Marco Valiente Rodríguez
DAW 1

# Índice

## Contenido

# 1. Introducción

En este informe, vamos a analizar el Test de Turing, el contexto de Alan Turing, su creador, sus implicaciones actuales, y cómo estas implicaciones siguen siendo relevantes hoy en día para la comprobación y teoría que las máquinas (and ergo, la inteligencia artificial), son capaces de llegar al nivel objetivo en donde pueden ser consideradas que estas son capaces de pensar, así como qué podría significar que las máquinas sean consideradas capaces de pensar y no de simplemente reconocer patrones y responder sin ser capaces de realmente entender de qué se está hablando o de qué se está comunicando.

# 2. Contextualización

Alan Turing fue un apasionado por la ciencia, y cuyos intereses que empezaron al principio siendo solo acerca de la química y la mente humana, se fue acercando más a la física y a las matemáticas. Su pasión por el desarrollo de las matemáticas le llevo a ámbitos lógicos y a la elaboración de algoritmos como un método definido de resolución de problemas. Con el proceso metódico analizado, empezó a desarrollar los términos de su investigación por términos de una máquina teórica capaz de hacer operaciones elementales, definidas, con símbolos y en una cinta de papel. Por esto, la contribución definitiva de Alan Turing fue la correspondencia entre operaciones lógicas, la acción de la mente humana, y una máquina que exista físicamente capaz de estas mismas. Esta relación evolucionó en el sistema moderno, de una máquina que escribiese las operaciones lógicas, a su capacidad de almacenar esos símbolos y poder usarlos como memoria, y planteó la posibilidad de qué tras esos, se pudiese elaborar una máquina (llamada Máquina de Turing Universal) que pudiese operar de forma autónoma y mejorar su propio código. Dichas teorías hacen que Alan Turing se volviese el padre teórico del ordenador y de la Inteligencia Artificial, dándonos muestra que estos avances tuvieron repercusiones históricas.

Turing, en 1950, con sus teorías de la máquina capaz de auto funcionar, se hizo la siguiente pregunta. ¿Pueden las máquinas realmente pensar? De esto, Turing planteó el siguiente experimento. Se pone el interrogador a hacer preguntas a un humano y a una máquina, y después se pone a intentar diferenciar cuál de los dos es cual. Este test teórico serviría para poder identificar qué nivel de inteligencia tendría una máquina emulando a la forma de pensar de un ser humano. Nunca fue especificado por Turing que el investigador podría estar hablando con una máquina, pero, de todas formas, la conclusión que Turing estaba haciendo es que una vez que se llegue al punto que la máquina y la persona no se puedan ser diferenciadas una de la otra, sería el punto en el que las máquinas serían consideradas capaces de pensar.

A día de hoy, ninguna máquina ha conseguido pasar el Test de Turing, la cual puede ser probada en los juegos para el premio Loebner. Todavía ni se encuentran cerca de ganar, a pesar de los avances. Aunque el Test de Turing no sea pasado, el premio sigue siendo entregado para aquellas máquinas que consiguen el nivel más cercano al de un ser humano. En 2018, el premio fue entregado al chatbot Kuki (o mitsuku), con su capacidad de emular coherentemente conversaciones humanas. Esta máquina usa reconocimiento de patrones y de estructuras para la formulación de respuestas a preguntas, incluso si el chatbot no tiene idea de que se está hablando, así como la creación de respuestas predefinidas. Otro caso famoso, es Eugene, que en 2014 llegó al precedente de un 33% de los jueces siendo engañados, aunque fue más mediático que científico. Desafortunadamente, falta de inversión ha hecho que el premio Loebner sea discontinuado.

En la actualidad, la inteligencia artificial está siendo un elemento esencial para el desarrollo moderno, incluyendo intentos de inversión por grandes empresas y su incorporación y normalización al día a día, para recopilación de material de prueba, así como su intento de mejora, con ChatGPT logrando uno de los mayores éxitos y fama en términos de desarrollo de inteligencia artificial y asistencia personal.

# 3. Documentación

Con el auge de ChatGPT, los modelos de inteligencia artificial han completamente cambiado el ámbito de comparación del Test de Turing, en el cual, por pruebas hechas a este, ha logrado una diferencia abrumadora no antes recordada.

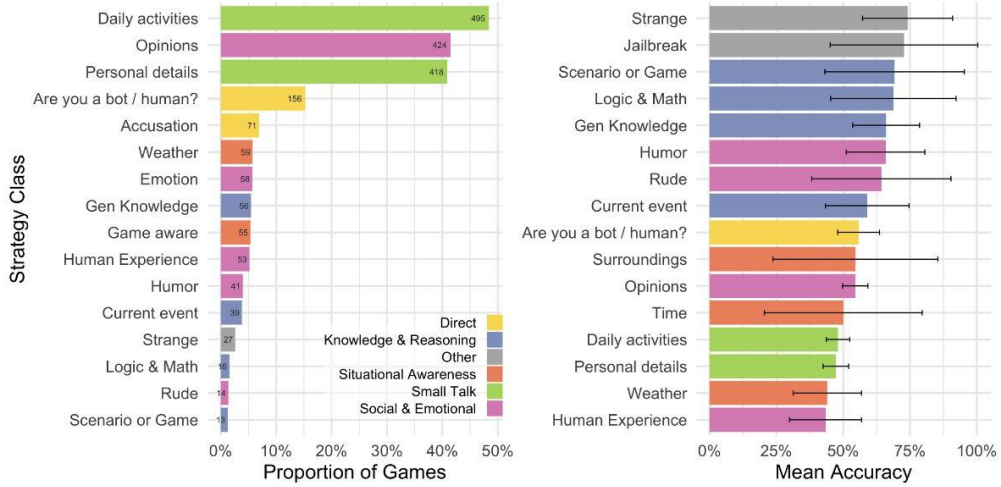



Figure 4: Classification of strategies employed by interrogators by proportion of games (left) and mean accuracy of games where strategies were deployed with 95% confidence intervals (right). Participants often engaged in small talk, asking witnesses about their personal details, activities, or opinions. Interrogators who said unusual things or used typical LLM "jailbreaks" were the most accurate.

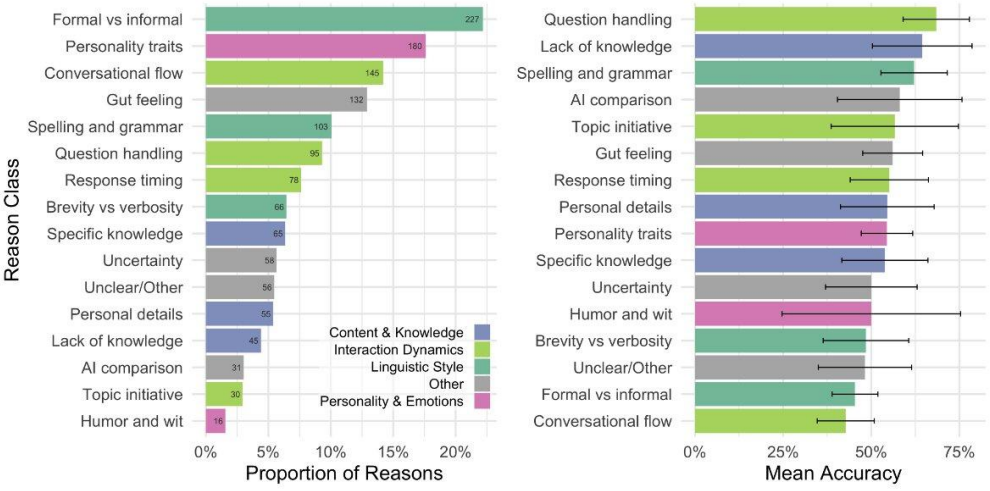



Figure 5: Proportion of interrogator reasons (left) and mean accuracy of verdicts that cited specific reasons with 95% confidence intervals (right). Interrogators were much more likely to cite linguistic style, conversational flow, and socio-emotional factors such as personality, rather than factors more traditionally associated with intelligence, such as knowledge and reasoning. The most accurate verdicts focussed on witnesses' directness in handling questions as well as instances where they lacked knowledge.

Con estos análisis, se pueden ver que la inteligencia artificial ha llegado a un nivel de respuesta que ya se considera que el Test de Turing puede ser superado, aunque solo cuando se le ordena a esta a 'disfrazarse' de una persona intentado adoptar esa personalidad, lo que ha llevado a ChatGPT-4.5 a tener un triunfo en este test, de un 73%. Ha marcado también un antes y un después, ya que, por ejemplo, las preguntas sobre temas matemáticos y estos tipos de conocimientos ya no son considerados como útiles para poder diferenciar a una máquina de una persona, pero en su lugar, el uso de preguntas más sociológicas y cotidianas, como se pueden mostrar en las dos tablas de arriba.

## A. Preparación

Para este experimento, en la universidad de San Diego, se tomaron un total de 300 participantes, que donde por 8 rondas, interactuaron en línea con dos interlocutores, uno siendo modelos como GPT-4.5, ELIZA, LLaMA 3.1-405B y GPT-4o, y el otro siendo un humano, eligiendo de los dos cuál era el más real. Los voluntarios actuaron como interrogadores a estas dos entidades, que eran consideradas 'testigos'.

## B. Interacción

Como podemos ver de los métodos anteriores, lo que triunfaron más para la evaluación de quien era el verdadero humano fue algo que se centró más en el análisis de cuáles de ellos eran capaces de responder a elementos cotidianos, como, por ejemplo, en las tablas se ven que se tomaron opiniones, actividades diarias, detalles personales, el clima, que, si sabían que estaban dentro del experimento o no, y este tipo de estrategias, por las masas de preguntantes.

## C. Decisión

Como se ha mencionado antes, el éxito de GPT-4.5 en superar el Test de Turing fue de un 73%, lo cual se desvía de la media hasta que el punto que las personas tenían más probabilidades de escoger al humano de forma correcta si escogiesen al azar, en contraposición con elegir por los métodos que los jueces usaron. Sin embargo, esto solo se debe a las instrucciones de intentar disfrazarse como una persona. Si en su lugar, no se les da esa instrucción, su éxito se desploma a sólo un 36%, lo que justifica que las máquinas no son capaces de poder todavía interactuar de forma autómata y con el razonamiento que superar el Test de Turing supondría, pero que han memorizado los comportamientos humanos que son capaces de catalogarlos y responder en los patrones de similitud adaptados.

Al mismo tiempo, esto confirma que la inteligencia artificial sí es capaz de engañar a humanos si esta misma es ordenada a ello.

# 4. Reflexiones

Estas pruebas y análisis son muestra pura de que hemos llegado a un punto de inflexión en nivel de las capacidades de las computadoras, que se han vuelto cada vez más buenas en copiar los elementos científicos y calificables de personas, como sus niveles matemáticos, y que están tomando pasos que sólo son de copia de patrones a la hora de tocar los temas sociales, en contraposición, con, por ejemplo, poder ser capaces de entender y elaborar con esas mismas prompts que los usuarios les preguntan a las máquinas.

Incluso con preguntas que fuese humanas y cercanas, los humanos fracasaron a la hora de poder determinar cuáles de ellos eran humanos y cuáles de ellos no. Por propia experiencia, con el uso de la página web que permite evaluar el Test de Turing tú mismo, me llegó a sorprender a la hora de hablar con un ChatBot que son capaces de incorporar no solo las formas formales e informales, que fue también un punto que los interrogantes usaron en el experimento de la universidad de San Diego, sino que son capaces de incorporar como jerga coloquial acotaciones como las palabras ‘lol’ que son comunes en el habla inglesa (usada para hablar con el bot), e incorporadas de manera natural, elementos que, por ejemplo, mayores son incapaces de hacer, aunque a pesar de eso, sí que se nota un cierto nivel de sospecha por el ritmo al que esa inteligencia artificial hablaba, que se puede ver en las tablas. Que los humanos podemos solo ver cómo hablar con una máquina puede sentirse ‘extraño’, para no mejor descripción.

Sin embargo, ¿significa esto que las máquinas pueden ser capaces de pensar ya, como postulaba Turing?

La respuesta sigue siendo, abiertamente para la mayoría de personas que pregunten, que no. Sin embargo, esto es sólo porque conocemos cómo funcionan las máquinas actuales, que no son autónomas. Los conocidos LLM (Learning Language Models). Sabemos que estas máquinas están diseñadas para evaluar el lenguaje humano e intentar adherirse para responder lo mejor posible que estas puedan, y no por usar los conceptos que estas sean capaces de razonar por su propia cuenta.

Superar este nivel es lo que se conoce como la diferencia entre los LLM y la ShortAI, y la verdadera revolución a la cual gobiernos de todo el mundo, incluidos desde China hasta Estados Unidos, han empezado a comprar y crear estructuras de bases de datos y de procesamiento electrónico para llegar, la GeneralAI, or GAI, que será considerada la verdadera Inteligencia Artificial, aquella que pueda mejorar su propio código, lo que teorizó Turing, que todavía no es posible con los niveles de GPT-5, pero que supone tener acceso a una inteligencia que puede hacer uno de los elementos que realmente califican a los seres humanos, la capacidad de aprender sin necesidad de crear un ser humano nuevo.

Que esto sea posible o no será el verdadero determinante de cuál será el impacto que la Inteligencia Artificial vaya a tener, y no un análisis de modelos que sean capaces de simplemente hacernos creer que son capaces de comprender.

# 5. Bibliografía


- https://blogs.uoc.edu/informatica/es/alan-turing-i-una-biografia-demasiado-corta-para-tan-gran-hombre/
- https://www.fundacionaquae.org/wiki/alan-turing-padre-la-inteligencia-artificial/
- https://www.studocu.com/pe/document/universidad-nacional-de-trujillo/inteligencia-artificial-1/informe-premio-loebner/102936325
- https://www.agenciasinc.es/Noticias/Una-maquina-pasa-por-primera-vez-el-test-de-Turing
- https://www.lesswrong.com/posts/mNR3tQDCxWL9XWv5u/on-the-loebner-silver-prize-a-turing-test
- https://quo.eldiario.es/tecnologia/q2504423718/chatgpt-ha-superado-el-test-de-turing
- https://humanornot.so/