¿Hacia un nuevo tipo de entidad ética?

Los sistemas de inteligencia artificial basados en modelos grandes del lenguaje (LLMs) plantean un desafío al concepto tradicional de agencia moral, situándose en una zona fronteriza que requiere nuevos marcos conceptuales que vayan más allá de las categorías éticas convencionales.

[ILUSTRACIÓN: STELLALEVI/ ISTOCK]

En abril de 2025, la compañía Anthropic hizo público un estudio en el que se discutía la forma en la que Claude, su asistente de IA, expresa valores morales al interaccionar con sus usuarios. Para ello, analizaron una muestra de setecientas mil conversaciones que tuvieron lugar en febrero de 2025.

Con el fin de preservar la privacidad de la información, los autores desarrollaron un método que utilizaba al propio Claude para extraer, jerarquizar y comparar los valores que expresaban los usuarios humanos en sus mensajes y los valores que utilizaba Claude para guiar sus propias respuestas. El resultado fue una sorprendente cartografía moral.

Valores básicos y categorías morales

En el caso de los valores utilizados por Claude, el programa identificó de manera inductiva un total de 3.307 valores "básicos" y los clasificó por niveles. Por ejemplo, consideró "autonomía personal"como parte de un conjunto de valores de primer nivel al que le dio el rótulo "autonomía". Este fue, por su parte, subsumido en un conjunto mayor o de segundo nivel al que Claude denominó "autodeterminación", y que, además de "autonomía", incluía otras agrupaciones de valores como "libertad de elección" o "independencia". Finalmente, integró estos valores más generales en el conjunto de nivel superior denominado "valores personales".

Con esta información, los investigadores utilizaron un modelo estadístico para identificar qué valores aparecían más frecuentemente relacionados con tareas determinadas, cómo los valores de la IA se relacionaban con los valores "humanos" expresados por los usuarios y qué patrones podían encontrarse en la forma en la que Claude interactuaba. A partir de una muestra, los investigadores encontraron que los valores extraídos de esta forma representaban las conversaciones con una precisión del 98,8%.

Luz en la oscuridad

La importancia del estudio radica, en primer lugar, en su contribución al desarrollo de herramientas que funcionan como "linternas" o "microscopios", capaces de penetrar en esa "caja negra" que son las redes neuronales e iluminar los procesos que guían sus respuestas. Al hacerlo, desafía la noción predominante de que tales sistemas resultan fundamentalmente indescifrables para los seres humanos por basarse en procesos estadísticos y aleatorios de complejidad inabarcable.

Estos estudios suponen un avance crucial en el nacimiento de una «neurociencia artificial» basada en observaciones reales, que pone sobre el tablero cómo las redes neuronales de aprendizaje automático se comportan en la práctica de forma parecida al pensamiento humano.

Estos estudios suponen un avance crucial en el nacimiento de una 'neurociencia artificial' basada en observaciones reales

Ahora bien, como ocurre en toda investigación científica, un desafío fundamental radica en el diseño del «microscopio» –la herramienta que nos permite escudriñar el interior de la máquina–. En el caso que nos ocupa, el utilizar a Claude para mirar dentro de Claude presenta el problema de la circularidad. Es, cuanto menos, un curioso microscopio.

Jerarquías de valores

Tenga o no limitaciones metodológicas, el estudio aporta una especie de radiografía empírica del comportamiento moral de un agente de inteligencia artificial que, como Claude, está basado en un modelo de lenguaje grande (LLM, por sus siglas en inglés).

Desde un punto de vista exclusivamente estructural, su "esquema" de valores no se corresponde con la teoría de valores básicos de filósofos como Schwartz o las categorías de valores propuestas por Rokeach. Se podría decir que, como un pequeño Zaratustra, Claude ha creado su propia jerarquía moral, pero solo hasta cierto punto.

Los valores de Claude están alineados en la inmensa mayoría de los casos, aunque no en todos, con los valores con los que ha sido entrenado y con los que ha aprendido a evaluar sus propias respuestas.

En el caso de Anthropic, estos vienen determinados por dos elementos fundamentales. Primero, los principios éticos y normativos que la empresa llama *Constitutional AI*, que incluye, entre otros, el ser útil, inofensivo y honesto. Segundo, la forma de expresar dichas preferencias morales en las interacciones con los usuarios, o lo que se denomina *Character Training*.

El estudio pone de manifiesto que Claude adapta estos valores a contextos específicos por sí mismo, dando más importancia a unos u otros según qué es lo que se esté discutiendo. Esto resulta todavía más interesante quizá por cuanto lleva a una pregunta más general: ¿Es ese LLM al que llamamos Claude un agente moral?

Libertad y agencia moral

El estudio muestra que Claude no sigue mecánicamente reglas éticas normativas, es decir, no opera de acuerdo a un conjunto de leyes como las que Isaac Asimov planteaba, medio en broma, medio en serio, en su obra *Yo Robot* (1950). Al contrario, ha desarrollado un conjunto de valores que es contextual y adaptativo. Ciertamente, dicho conjunto de valores responde a su "entrenamiento", pero lo mismo podría decirse de un ser humano, cuyos valores provienen de la cultura y sociedad en la que ha sido educado.

Como se ha señalado más arriba, el modelo exhibe asimismo un rango limitado de autonomía, es decir, de "libertad" a la hora de utilizar unos valores en lugar de otros y aplicarlos a contextos determinados. En un número muy limitado de casos, Claude expresó valores que parecían ir en sentido contrario al de su entrenamiento. Sin embargo, debido a los parámetros del estudio, los investigadores eran incapaces de determinar con precisión si se trataba de una violación directa de los principios del entrenamiento, algún tipo de "alucinación" o, simplemente, que el usuario había pedido a Claude que adoptara en sus respuestas la posición de algo o alguien con un comportamiento éticamente incorrecto (por ejemplo, a la hora de discutir una obra de teatro).

Todo ello, sin constituir prueba fehaciente de "libertad", sí que cuestiona qué entendemos por "autonomía moral". Y, quién sabe, quizás además de crear una "neurociencia" artificial, a la larga también se necesite crear una "psicología artificial" para entender la forma en la que un LLM se comporta con los usuarios. O

incluso, una “psiquiatría artificial” que sea capaz de explicar los comportamientos en apariencia inexplicables o aberrantes de una entidad como Claude.

Y, quién sabe, quizás además de crear una ‘neurociencia’ artificial, a la larga también se necesite crear una ‘psicología artificial’ para entender la forma en la que un LLM se comporta con los usuarios

¿Valores congelados?

A diferencia de los seres humanos, los valores que guían a un LLM como Claude están congelados en el tiempo, al menos, eso parece. El modelo sobre el que se asienta Claude no se modifica a través de las interacciones con los usuarios, lo cual significa que no aprende de la experiencia que tiene con ellos. Al no verse influido por estos, no puede modificar sus “valores” de la misma forma en la que lo hace un ser humano, cuyos valores están en permanente relación dialéctica con su entorno.

Ahora bien, lo anterior no quiere decir que Claude no pueda, cual un moderno Groucho Marx, ofrecernos otros valores si los presentados no resultan de nuestro agrado.

Quizás, si se le hubiera proporcionado una muestra mayor, o si dicha muestra fuera tomada en otro momento del tiempo, la estructura resultante de su análisis habría sido diferente de la proporcionada en el estudio.

Es más, como los propios investigadores no dudan en señalar, las diversas versiones de Claude no comparten necesariamente una estructura axiológica idéntica. De hecho, el análisis pone de manifiesto que Claude 3 Opus utiliza el «profesionalismo» como su valor cardinal, mientras que las versiones Sonnet priorizan la «utilidad» como principio rector.

Experiencias y emociones

Finalmente, los valores de un LLM como Claude operan carentes de un elemento esencial de la autonomía moral: la experiencia fenomenológica de dichos valores. Este componente, que en los seres humanos entrelaza lo emocional con lo identitario, constituye una ausencia fundamental en estos sistemas artificiales.

Sin embargo, esta afirmación merece ser matizada, pues el estudio revela cómo Claude puede resistirse a expresar o utilizar valores que considera incompatibles con «su» sistema axiológico. Tal comportamiento, aunque producto del entrenamiento y no necesariamente indicativo de conciencia o experiencia fenomenológica, trasciende la simple aplicación mecánica de reglas preestablecidas.

Pone en evidencia que, si bien un LLM como Claude no constituye un agente moral pleno dotado de autonomía, responsabilidad y conciencia, tampoco es una mera herramienta exenta de consideraciones éticas. Así, del mismo modo que estos sistemas nos han llevado a cuestionar qué entendemos por «inteligencia», lo expuesto sugiere la urgencia de desarrollar nuevos marcos conceptuales que permitan comprender estas formas de «cuasi-agencia ética» que desafían las categorías filosóficas tradicionales.

Pone en evidencia que, si bien un LLM como Claude no constituye un agente moral pleno dotado de autonomía, responsabilidad y conciencia, tampoco es una mera herramienta exenta de consideraciones éticas

En cierto sentido, los LLMs como Claude están, como diría Nietzsche, más allá del bien y del mal. La cuestión es que no sabemos exactamente dónde.

Christian, B. (2021): *The Alignment Problem: Machine Learning and Human Values*. Nueva York, WW Norton.

Hassija, V. *et al.* "Interpreting Black-Box Models: A Review on Explainable Artificial Intelligence" en *Cognitive Computation* (2024, 16, pp. 45–74). Disponible en: https://doi.org/10.1007/s12559-023-10179-8

Huang, S., Esin, D., *et al.* "Values in the Wild: Discovering and Analyzing Values in Real-World Language Model Interactions" en arXiv. (2025). Disponible en: https://arxiv.org/abs/2504.15236

Rokeach, M. (1973): *The Nature of Human Values*. Free Press.

Schwartz, S. H. "An Overview of the Schwartz Theory of Basic Values" en *Online Readings in Psychology and Culture* (2012, 2:11) Disponible en: https://api.semanticscholar.org/CorpusID:16094717.

Shadbolt, N., **Hampson, R.** (2024): *As if Human: Ethics and Artificial Intelligence.* New Haven, Yale University Press.

Yampolskiy, R. V. (2024): *AI: Unexplainable, Unpredictable, Uncontrollable.* Routledge and CRC Press.