MITOS Y REPRESENTACIONES DE LA INTELIGENCIA ARTIFICIAL

Editores

Gastón Becerra | Joaquín Mezzadra | Guillermo Movia

UFLO
UNIVERSIDAD

“La IA es objetiva”. Aportes de la filosofía de la ciencia para una discusión social

Gastón Becerra y Joaquín Mezzadra

La IA en la grieta

El 9 de febrero del 2023 el diputado argentino Rodrigo De Loredo (Unión Cívica Radical) leyó el siguiente texto en una comisión de la Cámara: “Este texto, presidenta, no fue escrito por mí, ni tampoco por nadie que integre mi equipo, ni por ningún diputado de mi bloque, ni es un texto que haya sido escrito por juristas y politólogos, que tenemos en abundancia en la Argentina. Este es un texto que fue escrito recién por un chat de inteligencia artificial, GPT-3, que tiene un modelo de 175 millones de parámetros para construirse, y que contestó la pregunta ‘¿Por qué los populismos en el mundo tienden a controlar los poderes judiciales de sus Estados?’”. El contexto de su intervención era el del tratamiento legislativo de un pedido de juicio político a la Corte Suprema de Justicia impulsado por el bloque opositor, comúnmente denominado “kirchnerismo”, en relación a los ex presidentes Néstor Kirchner (2003-2007) y Cristina Fernández de Kirchner (2007-2015). De Loredo quería así justificar que el proyecto legislativo no era más que un intento de intervención populista.

Del otro lado de la grieta política, el 14 de noviembre de 2024, la ex presidenta Cristina Fernández de Kirchner publicó un *tweet* para criticar un fallo judicial de la Cámara de Casación en el marco de una causa por administración fraudulenta en su contra, señalando que había pedido a ChatGPT analizar el fallo con “una visión objetiva”, y que habría encontrado inconsistencias. Luego, desafió al actual presidente Javier Milei en estos términos: “Sé que a vos te gusta la inteligencia artificial (a mí también me gusta mucho) y ya que en unos días te vas a EEUU y lo vas a ver a Elon Musk, ¿por qué no le preguntas cómo podemos hacer para crear un Poder Judicial con inteligencia artificial? ¿Te imaginás la guita que se ahorrarían el Estado y todos los argentinos?”.

En el primer ejemplo, el diputado De Loredo parece creer que ChatGPT puede cerrar una controversia histórica y política, es decir, dar una demostración sin prejuicios que se pueda tomar como verdad. En el segundo, la ex presidenta Fernández de Kirchner refiere a ChatGPT como imparcial o libre de intereses, y en última instancia, justo. Ambos, de alguna manera, plantean una separación entre lo subjetivo y lo objetivo, enfrentando a los humanos (juristas, jueces, historiadores, políticos) con la IA. Sin embargo, en ciencia sabemos que la objetividad es algo difícil

de alcanzar, o que incluso hay muchas definiciones de objetividad. Y si bien es muy discutible que la objetividad científica sea equivalente a la jurídica, la periodística, la política, o a la manera en que se entienda la objetividad en otros ámbitos, algunos elementos de la reflexión científica podrían ayudarnos a plantear preguntas para el tratamiento de este mito.

De qué objetividad estamos hablando

Primero, aclaremos qué entendemos por objetividad. Una idea común es que algo es objetivo si refleja el mundo tal como es, sin interpretaciones. Sin embargo, vamos a dejar de lado esa definición, ya que plantea preguntas filosóficas complicadas sobre lo que realmente existe y lo que podemos conocer.

Otra definición de objetividad dice que se logra cuando se es "valorativamente neutral", es decir, cuando nuestros valores, creencias y preferencias sociales no afectan el trabajo científico. Pero, ¿puede la IA ser realmente neutral en ese sentido?

Varios investigadores señalan que la IA, especialmente la IA generativa, puede tener "sesgos", es decir, inclinaciones sistemáticas presentes en los modelos y algoritmos que favorecen o desfavorecen de manera consistente a ciertos grupos sociales, reforzando estereotipos, provocando discriminación, o dando lugar a decisiones injustas. Los sesgos pueden tener origen a lo largo de toda la cadena de decisiones involucradas en el desarrollo de un sistema. El tipo de sesgo más conocido y documentado es el que refleja en los datos las disparidades e injusticias sociales existentes: por ejemplo, cuando un modelo de selección de personal es entrenado con los CVs de los ejecutivos de una empresa y termina por reproducir un cierto "techo de cristal" para las mujeres, ya que generaliza las características del grupo conformado generalmente por varones. De igual modo, puede ser que la base de entrenamientos no incluya la misma cantidad y calidad de datos para todas las subpoblaciones. Así, por ejemplo, se ha denunciado que los algoritmos de reconocimiento facial fallan en reconocer rostros afroamericanos por estar entrenados mayormente con fotos de personas caucásicas. También se introducen sesgos en los momentos de evaluación y programación del modelo, en las políticas de los equipos de control, entre otros. En suma, los sistemas pueden reflejar valores y preferencias de sus fuentes de datos y de sus desarrolladores, lo cual constituye un obstáculo para cualquier pretensión de una IA objetiva, y mucho menos, justa.

No obstante, el problema de los sesgos en la ciencia no es novedoso. Gran parte de la filosofía de la ciencia del siglo XX ha mostrado que distintos valores sociales tienen un rol epistémico importante al guiar las preguntas y la formulación de problemas de investigación. Entonces, se ha buscado ensayar una objetividad que

reconozca la existencia de valores en la ciencia e igualmente se pregunte cómo lograr conocimiento objetivo. Esto es lo que propone el sociólogo francés Pierre Bourdieu (2001) o la epistemóloga feminista norteamericana Helen Longino (1990). El primero entiende que la ciencia es un campo de lucha entre actores con intereses contrapuestos y que la objetividad reside entonces en el acuerdo de reglas que regulen su competencia y en la reflexión constante sobre estos condicionamientos sociales. La propuesta de Longino consiste en que la objetividad puede conseguirse si se cumplen ciertos requisitos: se deben adoptar canales de diálogo aceptados por todos, respetar la igualdad de estatus o autoridad científica entre los partícipes, establecer criterios y estándares de crítica y obtención de conocimiento compartidos, así como una actitud atenta y reactiva de la comunidad.

Así, la objetividad no es entonces un desafío que se consiga adoptando una actitud o un estado de la mente, sino más bien una dinámica, una forma organizada de interacción. El trayecto de la objetividad en la ciencia nos sugiere que, contrario a lo que proponían De Loredo o Fernández de Kirchner, una respuesta objetiva no es la que no deja espacio para la subjetividad, sino la que confía en mecanismos intersubjetivos, es decir, que es producto de la discusión y los posibles acuerdos entre sujetos que se hacen cargo de sus valores y subjetividades y que aceptan revisar y tematizar sus sesgos. Tal vez, entonces, para poder aprovechar la enorme potencia de la IA, la clave no sea buscar cerrar los debates sino generar instancias de comunicación y crítica más sólidas, que incluyan también a la IA en esta comunicación. En cualquier caso, la idea de que la tecnología es neutral, el mandato de que debemos utilizar la tecnología para evitar conflictos sociales y cerrar desacuerdos parece obturar el debate.

***

En la escritura de este breve análisis no utilizamos herramientas de IA, sin embargo, por curiosidad y en ánimos de una comparación, le pedimos a ChatGPT 4o-mini que escriba un breve análisis de este mito siguiendo el *template* del ejercicio. En la introducción, la voz que destacó no fue la de políticos argentinos sino la de "empresarios tecnológicos como Elon Musk y Mark Zuckerberg". En los argumentos no diferimos mucho: para ChatGPT el *quid* de la cuestión también reside en si la IA está "impregnada de las subjetividades de aquellos que las crean y entrenan", algo que nos llevó a los problemas de los sesgos. Finalmente, advirtió sobre los riesgos de la automatización y el reemplazo; aunque, sin una teoría "sustantiva" o una definición de la objetividad que pudiera abrir nuevas dimensiones, se quedó en comentarios éticos generales. Más interesante fue la conversación cuando le repetimos la pregunta de De Loredo acerca del populismo y le fuimos dando contexto para, en última instancia, corroborar si ChatGPT cree que el kirchnerismo

es populista. Parece que al final, el que fue injusto fue De Loredo, que le ocultó a la aplicación información sensible sobre los juicios que buscaba.

Referencias

Bourdieu, P. (2001). *El oficio del científico*. Anagrama.

Longino, H. (1990). *Science as social knowledge. Values and objectivity in scientific inquiry*. Princeton University Press.