Ser éticos con la inteligencia artificial

Juan Ignacio Rouyet

¿Es la inteligencia artificial la que tiene que ser ética o el ser humano cuando la utiliza? He aquí las claves para entender el uso ético de la IA.

Una inteligencia artificial ni tan buena ni tan mala

En 1889 el periódico inglés *Spectator* alertaba sobre los dañinos efectos intelectuales de la electricidad, y, en particular, del telégrafo. Declaraba que el telégrafo era un invento que claramente iba a afectar al cerebro y al comportamiento humano. Debido a este nuevo medio de comunicación, la población no paraba de recibir información de modo continuo y con muy poco tiempo para la reflexión. El resultado llevaría a debilitar y finalmente paralizar el poder reflexivo.

Nicholas Carr se preguntó algo similar respecto a Google. Ciertamente, viendo las patochadas que a veces se publican en las redes sociales, uno pudiera pensar que tal debilidad de reflexión ya se ha hecho realidad. Sin embargo, de ser verdad, no sería justo atribuirlo al desarrollo de la electricidad, sino más bien a la propia naturaleza de los hacedores de tales despropósitos.

Pero no todo iban a ser desgracias. El telégrafo también estaba llamado a traer la paz al mundo, al más puro deseo de un certamen de belleza. El telégrafo podía transmitir la información a la velocidad del rayo. Esto iba a permitir favorecer la comunicación entre toda la humanidad de forma instantánea, lo cual nos llevaría a una conciencia universal que acabaría con la barbarie. No parece que haya sido así. Hoy gozamos de una capacidad de comunicación inmediata sin precedentes, que permite compartir sin tardanza las imágenes de tu noticiable desayuno. Sin embargo, la barbarie sigue siendo nuestra compañera en la historia.

Algo similar sucede hoy en día con la inteligencia artificial: ni tan buena, ni tan mala. La inteligencia artificial trabaja principalmente identificando patrones y en esto es en verdad excelente. La cuestión radica a qué dedicamos dicha excelencia.

Esta identificación de patrones, permite, por ejemplo, identificar síntomas de COVID-19 en personas asintomáticas, evitando así la extensión de la enfermedad; también puede reconocer y clasificar leopardos por las manchas de su pelaje, favoreciendo su seguimiento y disminuir el riesgo de extinción; o bien reconocer trazos en la escritura de textos antiguos y determinar su número de autores. Es la cara amable y deseada de la inteligencia artificial que nos ayuda a mejorar la salud, el medio ambiente y la ciencia.

Esta misma grandeza en reconocer patrones es la que permite a la inteligencia artificial clasificarnos (encasillarnos), por nuestra actividad en el uso de aplicaciones móviles o en redes sociales, para predecir nuestro comportamiento e incitarnos a comprar productos sin apenas ser conscientes. Nos puede recomendar aquello que realmente sabe que nos gusta o que necesitamos. La inteligencia artificial sabe más de nuestra intimidad que nosotros mismos. Este conocimiento sobre nuestra forma de ser es el que permite también generar notificaciones, cuidadosamente pensadas para mantenernos activos en dichas redes sociales, llegando a la adicción, tal y como denuncia el documental *The Social Dilema*.

Es un conocimiento matemático sobre nuestra personalidad que niega a la libertad una opción de existir. El año pasado, durante los meses más duros de la pandemia que suspendieron las clases docentes y la celebración de exámenes, en el Reino Unido se decidió utilizar un sistema de inteligencia artificial para determinar la evaluación académica de ese curso, en función de las calificaciones obtenidas en años anteriores. ¿Acaso no puedo ser distinto este año respecto al anterior y ahora estudiar más? ¿No es posible que mis actos futuros sean diferentes a mis patrones de comportamiento del pasado? ¿No aceptas que soy libre? Para la inteligencia artificial, la respuesta a todas estas preguntas es “no, no puedes, así lo dice mi algoritmo”.

Éticas para una inteligencia artificial

Estas preguntas en el fondo nos están hablando sobre la moral, aunque este término nos parezca algo “viejuno” (neologismo despectivo instagramer para decir antiguo, bien distinto, dicho sea de paso, de “vintage”, concepto elogioso de la misma antigüedad). Así es, porque la moral nos interpela sobre cómo actuar y sobre cómo responder, ante nosotros y ante la sociedad, sobre tales actos. Al hablar de actuar podemos pensar en nuestras actuaciones directas, o en aquellas a través de una herramienta como es la inteligencia artificial. Y si queremos pensar sobre la moral, debemos recurrir a la ética, que es la parte de la filosofía que se ocupa de la misma. ¡Quién nos iba a decir que la filosofía podía servir para algo práctico en nuestra vida! Pues sí, detrás de nuestra forma de pensar y de proponer soluciones se encuentran propuestas filosóficas.

Ese lado no tan amable de la inteligencia artificial que acabamos de ver es materia de preocupación. Como solución a ello existen ciertas propuestas que en el fondo se alimentan de otras filosóficas.

Una de las primeras soluciones a los dilemas éticos de la inteligencia artificial son las famosas Tres Leyes de la Robótica de Asimov. Constituyen tres principios éticos que supuestamente se

deberían programar en un sistema inteligente. La primera de ellas, por ejemplo, dice que un robot no hará daño a un ser humano o, por inacción, permitirá que un ser humano sufra daño. Detrás de esta solución práctica se encuentra una visión de la ética de Kant, según la cual debemos actuar por un principio categórico, no condicionado por nada más. Tales leyes de la robótica serían esos principios categóricos de los sistemas inteligentes. Pero, como buena ética kantiana, tiene sus dificultades prácticas.

El primer problema surge de tratar con conceptos abstractos, tales como “hacer daño”. Matar es claramente hacer daño; poner una vacuna también causa dolor. ¿Hasta dónde llega el daño? El segundo problema viene de la evaluación del posible daño y de tener que evitar éste a todo ser humano. Si un sistema inteligente se encuentra con dos personas con la misma probabilidad de daño, no sabrá a quien atender y puede ocurrir que se acabe bloqueando.

Otra solución al comportamiento ético de la inteligencia artificial podría ser recurrir a una visión de la mayoría. Sí, ha leído bien: la mayoría decide qué está bien y qué está mal. Por ejemplo, supongamos que tenemos que decidir qué debe hacer un vehículo de conducción autónoma en el supuesto de accidente inmediato e irreversible que implica a viandantes: ¿salvamos a los ocupantes o a los viandantes? El Instituto de Tecnología de Massachusetts (MIT) propone su Moral Machine para analizar estas posibles decisiones. Esta solución se ampara en la llamada ética utilitarista de Bentham según la cual una acción es buena si proporciona la mayor felicidad para el mayor número.

Este tipo de soluciones éticas se preocupan por las consecuencias de las acciones y no tanto por los principios que las inspiran, siendo la felicidad una de las principales consecuencias deseadas. Desde esta perspectiva utilitarista, las recomendaciones en redes sociales para ver nuevos contenidos serían éticamente aceptables. El sistema simplemente nos recomienda lo que más le satisface a la mayoría, aquello que más “*likes*” ha conseguido. ¿Y si tanta recomendación me causa adicción, como hemos visto? Eso ya es responsabilidad de cada uno, se podría argumentar, pero el fin es bueno: las recomendaciones buscan mejorar la experiencia del usuario.

Principios, consecuencias y responsabilidad

En la última frase del párrafo anterior se encuentra la tercera variable de la solución ética a la inteligencia artificial: es responsabilidad de cada uno. Hemos visto que la moral consiste en la toma de decisiones y en cómo respondemos, ante nosotros y la sociedad, de dichas decisiones. El comportamiento ético de un sistema inteligente siempre debe concluir en la responsabilidad de una persona (o institución): de la persona responsable de la conducción de un vehículo autónomo; de la persona que determina una sentencia ayudado por un sistema inteligente; o de la persona que contrata en función de recomendaciones de *machine learning*. Esta responsabilidad es una dimensión más de una inteligencia artificial ética, pero no la única.

Existen además aquellas otras dos dimensiones que hemos visto en los ejemplos de posibles éticas a aplicar. Hemos hablado de una ética de principios (por ejemplo, Kant) y de una ética de consecuencias (por ejemplo, el utilitarismo). La solución ética a los sistemas inteligentes consiste en tres elementos: primero, determinar los principios que rigen dicho sistema; segundo, evaluar sus consecuencias y tercero, permitir que las personas afectadas puedan tomar decisiones.

Marco europeo para una inteligencia artificial fiable

Ésta es la idea que inspira a las propuestas de la Unión Europea en dos de sus textos principales para una inteligencia artificial fiable. En el documento *Directrices Éticas para una Inteligencia Artificial Fiable* establece el concepto de inteligencia artificial fiable como aquella que es lícita (que cumple con la normativa vigente), robusta (sin fallos para no causar daños) y ética (que asegure el cumplimiento de valores éticos). Sobre esta visión se construyen las tres dimensiones que comentamos.

Para que una inteligencia artificial sea fiable ésta debe partir de unos principios, que para la UE son: prevención del daño, equidad, explicación y autonomía humana, siendo este último el que garantiza esa capacidad de decisión. La inteligencia artificial no puede subordinar, coaccionar, engañar, manipular, condicionar o dirigir a las personas de manera injustificada. Al contrario, la inteligencia artificial debe potenciar las aptitudes cognitivas, sociales y culturales de las personas. Ésta es la dimensión de responsabilidad.

Para ayudar al cumplimiento de estos principios y garantizar esta responsabilidad, el marco de directrices establece una serie de requisitos claves a evaluar en cada sistema inteligente, tales como la acción y supervisión humana, la solidez técnica, la privacidad, la transparencia o la no discriminación. Estos requisitos clave se orientan a prevenir esas posibles consecuencias.

Para que todo esto no se quede en palabras bonitas, al albur de la buena fe y dado que la mejor forma de animar a dicha buena fe es la aplicación de medios coercitivos, la Comisión Europea ha propuesto una legislación con medidas concretas y sanciones sobre el uso de sistemas inteligentes. Esta normativa prohíbe ciertos usos de la inteligencia artificial, como, por ejemplo, aquellos que manipulan el comportamiento para eludir la voluntad de los usuarios o sistemas que permitan la “puntuación social” por parte de los Gobiernos. Otros usos de la inteligencia artificial, tales como su aplicación en procesos legales, control de migraciones, usos educativos o gestión de trabajadores, son considerados como de alto riesgo y deben tener un proceso previo de auditoría para su evaluación de conformidad antes de ser introducidos en el mercado. Este tipo de auditorías es demandado por distintas organizaciones, como We The Humans, *think tank* orientado hacia una inteligencia artificial ética.

El telégrafo no parece que nos haya hecho más tontos, ni ha eliminado la barbarie. Es un error pensar que una tecnología por sí sola va a resolver, o va a ser la causa de todos nuestros problemas. Todo depende del uso que hagamos de ella. Depende de nosotros, de nuestra responsabilidad, que la inteligencia artificial tenga un uso ético. La pregunta no es cómo hacer que la inteligencia artificial sea ética, sino cómo hacer que nosotros seamos éticos usando la inteligencia artificial.

Fecha de creación
26 mayo, 2021