ÉTICA APLICADA EN LA INTELIGENCIA ARTIFICIAL

De robots éticos a personas éticas con robots

JUAN IGNACIO ROUYET

Pensamos que la inteligencia artificial es autónoma, pero tan solo es un objeto sujeto a un *software*. Por ello quizá no quepa hablar de una ética para las máquinas sino de una ética para los humanos que utilizan la inteligencia artificial. ¿Con qué objetivo? Buscar la vida buena aristotélica. ¿Con qué ética? Ése es el debate.

Palabras clave:
Inteligencia artificial, ética aplicada, ética, autómata, computación, vehículos autónomos.

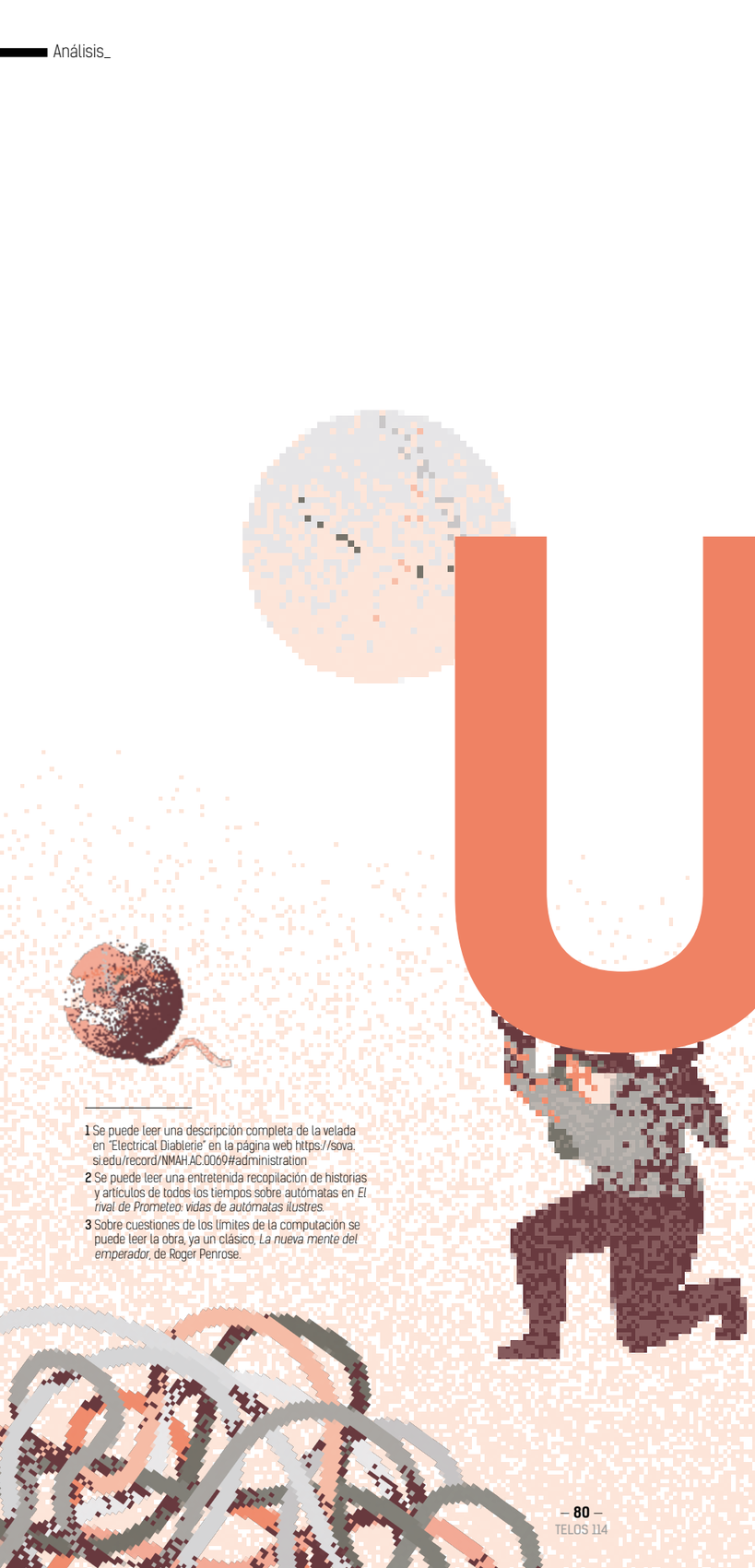
Ethics applied to artificial intelligence

FROM ETHICAL ROBOTS TO ETHICAL PEOPLE WITH ROBOTS

We think that artificial intelligence is autonomous, but it is only an object subject to software. So perhaps it is not possible to speak of ethics for machines, but of ethics for humans who use artificial intelligence. For what purpose? Seeking the good Aristotelian life. With what ethics? Thi is the debate.

Keywords: *Artificial intelligence, applied ethics, ethics, automaton, computing, autonomous vehicles.*

ILUSTRACIÓN: **SANDRA RILOVA**



“Un robot ha presidido nuestra cena de fin de año”. Con este pensamiento concluyeron los comensales una Nochevieja de 1884. Habían sido invitados por William J. Hammer, antiguo ayudante de laboratorio de Edison, a una amena y sorprendente “cena eléctrica”. En la sala donde se celebró la velada, Hammer aparejó una gran mesa alargada, sobre la cual dispuso cuidadosamente un “electrificante” menú, compuesto, entre otras delicias, por “tostada eléctrica”, “pastel de telégrafo”, “pastel de teléfono” o “limonada incandescente”. La mesa estaba presidida en su extremo por un autómata llamado Júpiter. A las 12 en punto de aquella noche, la luz se apagó y distintos elementos de la sala se fueron encendiendo. Entre fogonazos eléctricos, el pastel de telégrafo comenzó a emitir mensajes y la limonada incandescente se iluminó; Júpiter levantó su copa y empezó a beber, sus ojos brillaron con un verde intenso, su nariz enrojeció, en su pecho brillaron luces diamantinas y con voz profunda y jocosa empezó a gritar: ¡Feliz año nuevo! ¡Feliz año nuevo! Al finalizar la velada los invitados de Hammer partieron con la inquietante sensación de haber vivido acontecimientos con medio siglo de antelación[1].

1 Se puede leer una descripción completa de la velada en “Electrical Diablerie” en la página web https://sova.si.edu/record/NMAH.AC.0069#administration

2 Se puede leer una entretenida recopilación de historias y artículos de todos los tiempos sobre autómatas en *El rival de Prometeo: vidas de autómatas ilustres*.

3 Sobre cuestiones de los límites de la computación se puede leer la obra, ya un clásico, *La nueva mente del emperador*, de Roger Penrose.

La inteligencia artificial solo simula autonomía. Un sistema inteligente ajusta sus acciones según el entorno para conseguir un objetivo dado

Hoy en día esta “cena eléctrica” y el propio robot Júpiter no tienen misterio para nosotros[2]. Todo ese aparato eléctrico no era más que un conjunto de artilugios electromecánicos operados por Hammer mediante una serie de interruptores controlados desde un cuadro de mandos que descansaba en su regazo. Júpiter era capaz de hablar porque disponía de un fonógrafo ubicado en el interior de su cuerpo, accionado también por Hammer. Todo el invento estaba alimentado por unas baterías colocadas debajo de la mesa. ¿Podemos afirmar que era un sistema inteligente?

Depende de lo que entendamos por inteligencia y de lo que incluyamos dentro del sistema. De manera simplificada podemos asimilar por inteligencia la capacidad de pensar y actuar de manera racional como un ser humano. Si por sistema consideramos solo al autómata Júpiter, no podemos decir que exista comportamiento racional, pues todo él estaba accionado por Hammer. Por el contrario, si por sistema entendemos todo lo anterior junto al propio señor Hammer, entonces no tendremos duda en admitir que estamos delante de un sistema inteligente (considerando al señor Hammer racional, a pesar de su locura de cena).

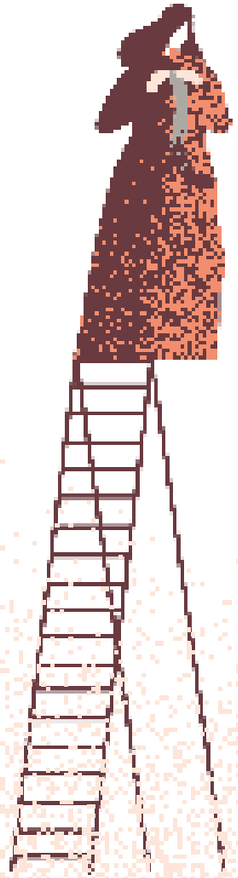
El autómata Júpiter es un rudimento aproximado pero válido de lo que hoy entendemos por inteligencia artificial. ¡Qué dislate!, se podrá pensar. Júpiter no tomaba decisiones. La inteligencia artificial actual tampoco; sus decisiones están condicionadas por un *software* desarrollado por unas personas. La confusión viene de creer que la inteligencia artificial es autónoma y nos ilusionamos hablando de vehículos de conducción autónoma. Sin embargo, en estos vehículos, quien se encuentra al volante es un, o una, ingeniero, a quien no conocemos, que toma sus decisiones sobre qué interruptor activar para, por ejemplo, en caso de accidente salvar a éste o aquél. En lugar de vehículos de conducción autónoma deberíamos llamarlos vehículos de conducción desconocida. Al menos en la “cena electrificante”, Hammer estaba en la mesa con sus invitados y estos le conocían.

La inteligencia artificial solo simula autonomía. Un sistema inteligente ajusta sus acciones según el entorno para conseguir un objetivo dado. Este ajuste lo realiza en un proceso de prueba y error llamado “aprendizaje”, el cual, junto con sus acciones de adaptación al entorno, simulan una ilusión de autonomía. Una ilusión, pues tan solo es el resultado de un *software* que le hace actuar según la intención de su desarrollador, de igual manera que Júpiter se movía según sus piezas mecánicas activadas por Hammer. La inteligencia artificial no es un sujeto, sino un objeto sujeto a un *software*.

Dado que un sistema de inteligencia artificial está sujeto a un *software*, deberemos crear un *software* ético. Pero ¿con qué ética? Y, si encontramos una ética adecuada, ¿será ésta computable?[3]

Con la primera pregunta llevamos 2.500 años y no hemos llegado a una solución concluyente. Una agrupación, que no la única, de los tipos de éticas que se han sucedido a lo largo de la historia divide a éstas en dos categorías: éticas teleológicas (o de las consecuencias) y éticas deontológicas (o de los principios).

Las éticas teleológicas determinan que una acción es correcta en función de su resultado o consecuencia. Así para Aristóteles, una acción es buena si consigue la felicidad; o para los utilitaristas, si se consigue el mayor bienestar para el mayor número. Ahora bien, ¿obtener el mayor bien para muchos es lo que se debe hacer? Con esta pregunta entran en juego las éticas deontológicas, donde lo correcto viene determinado por el cumplimiento del deber, con independencia de sus consecuen- ➠



cias. Aquí destaca Kant, con su "imperativo categórico", que podemos asimilar a un "mandato incondicionado".

En un principio, ambas éticas pueden ser computables. Las más sencillas de programar serían las deontológicas. Bastaría con incluir estos imperativos categóricos como órdenes expresas para que el sistema inteligente realice u omita una acción. Pero, ¿qué mandato programamos? ¿Sería universal o puede depender del usuario? Si queremos cumplir con una ética teleológica, el sistema debería hacer una predicción sobre las consecuencias de sus actos, para lo cual tendría que plantearse varias acciones posibles y hacer un cálculo estadístico y predictivo de la probabilidad de bondad o beneficio de cada consecuencia, actuando entonces con la acción de beneficio probable más alto. Esto reduce la ética a un cálculo matemático. Entonces, ¿cómo calculamos la bondad o beneficio de una acción? ¿Es la ética una cuestión de estadística? Si finalmente no sucede el beneficio más probable, ¿quién responde?

Tener una inteligencia artificial ética es tener seres humanos que buscan ser mejores personas usando la inteligencia artificial

Afortunadamente hay una posible solución a este círculo filosófico entre las éticas deontológicas y las teleológicas. La solución está en la llamada ética aplicada, que consiste en circular entre la ética de principios y la ética de las consecuencias con la mediación de las virtudes.

Para Aristóteles, la virtud consiste en realizar bien su función. Así, un ser humano virtuoso sería aquel que realiza bien su función ¿Y cuál es mi función como ser humano? Entramos de nuevo en siglos de debate. Actualmente hablamos de virtud en el sentido de la excelencia en la persona que busca un comportamiento moral[4], es decir, que busca la vida buena. En palabras de Alasdair MacIntyre[5], la vida buena para el hombre es la vida dedicada a buscar la vida buena para el hombre, y las virtudes nos capacitan para entender más y mejor lo que es la vida buena para el hombre.

De todas estas cuestiones filosóficas extraemos dos conclusiones relevantes para una computación de la ética. Primero, que esto mismo resulta complicado. Un código ético computable debería tener éticas deontológicas, éticas teleológicas y virtudes: las dos primeras podrían ser computables, como hemos visto, pero veo complejo cómo convertir la virtud en un algoritmo. Por consiguiente, y ésta es la segunda conclusión, la única salida para disponer de un sistema inteligente ético no es tanto computar un código ético, cuestión ardua, sino considerar al ser humano dentro de dicho sistema —como el sistema formado por el autómata Júpiter y su hacedor Hammer, donde la ética de Júpiter es la ética de Hammer—. De esta manera, tener una inteligencia artificial ética es tener seres humanos que buscan ser mejores personas usando la inteligencia artificial mediante la ética aplicada.

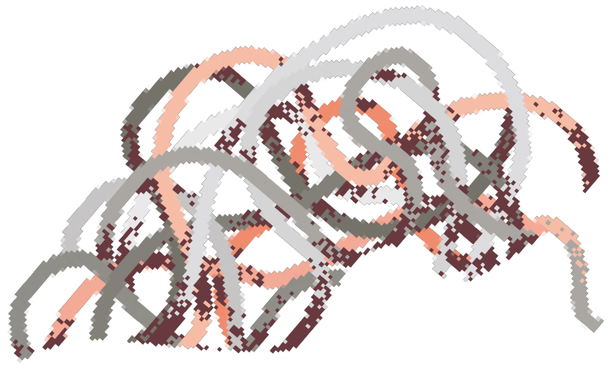
La ética aplicada intenta resolver problemas éticos de actividades humanas concretas. En este sentido ha sido el modelo para crear marcos éticos como la bioética, ética de la economía o ética de las profesiones. Siguiendo a Adela Cortina[6] proponemos usar la ética aplicada mediante este método circular —llamado hermenéutico— entre la ética de los principios y la ética de las consecuencias, con una mediación de las virtudes, de la siguiente forma:

- Determinar el fin específico —o bien interno— por el que cobra sentido y legitimidad social la inteligencia artificial.
- Esclarecer los medios que usa la inteligencia artificial para producir dicho bien en la sociedad.
- Indagar qué virtudes, valores y principios debemos incorporar para alcanzar ese bien interno, dentro de una moral cívica de la sociedad en la que se inscribe y mediante lo que se llama la ética del discurso.
- Dejar la toma de decisión en manos de los afectados, los cuales, con asesoría y con datos precisos y claros, puedan ponderar las consecuencias, sirviéndose de criterios tomados de distintas éticas —una de ellas podría ser la utilitarista—.

Por tanto, la ética en la inteligencia artificial no es cuestión —solo— de emitir códigos de buenas prácticas por parte de las organizaciones (códigos deontológicos), sino de profundizar en cuál es el fin específico de la inteligencia artificial, qué virtudes queremos desarrollar para conseguir tales fines y cuáles son sus consecuencias. Sobre este último punto todavía necesitamos más investigación. Para los dos primeros, lanzo una propuesta inicial. La inteligencia artificial es una herramienta, como lo es una palanca o un martillo, por tanto, su fin es aumentar las capacidades del ser humano; el fin de la inteligencia artificial es ayudar al ser humano.

Para conseguir este fin, una de las virtudes que debemos aplicar es la autonomía, que consiste en obedecer a esa parte de cada uno que es libre porque está sujeta a la razón. Así, un sistema inteligente dejaría de ser ético si usurpa dicha autonomía y evita que nosotros tomemos decisiones. Puede sonar algo brusco, pero la decisión de atropellar a alguien o estrellar el coche debe seguir siendo nuestra, porque eso es una decisión del ámbito de la ética y la ética es algo específicamente humano. Para tomar la decisión correcta tenemos la ética aplicada.

En la cena de fin de año de 1884 hubo un sistema inteligente formado por el autómata Júpiter y por Hammer, un ser humano autónomo. Esta idea nunca debemos perderla.



Bibliografía

Camps, V. (2013): *Breve historia de la ética.* Barcelona, RBA.

Cortina, A. y Martínez, E. (2001): *Ética.* Madrid, Akal.

Ferrater, J. y Cohn, P. (1981): *Ética aplicada. Del aborto a la violencia.* Madrid, Alianza Universidad.

MacIntyre, A. (1987): *Tras la virtud.* Barcelona, Crítica.

Penrose, R. (1991): *La nueva mente del emperador.* Barcelona, Penguin Random House.

VV. AA. (2009): *El rival de Prometeo: vidas de autómatas ilustres.* Madrid, Impedimenta.

4 Camps (2013), p. 398.
5 MacIntyre (1987), p. 271.
6 Cortina (2001), p. 165.