EL BASILISCO

Revista de materialismo filosófico

Nº 59 (2023), páginas 44-48

Gonzalo Génova Fuster

Universidad Carlos III de Madrid

Inteligencia artificial: explicabilidad, racionalidad y responsabilidad profesional del ingeniero

Resumen:

El autor pretende responder a la siguiente pregunta: ¿qué tipo de responsabilidad puede asumir un ingeniero que utiliza herramientas de inteligencia artificial para desarrollar su trabajo? Obviamente, los ingenieros siempre han utilizado todo tipo de herramientas: conceptuales, matemáticas, mecánicas. Entonces, ¿qué añaden las técnicas de inteligencia artificial que pueda suponer un problema adicional? ¿Por qué esto es un problema nuevo, que requiere una reflexión específica? Por decirlo muy brevemente, el problema sería este: ¿cómo voy a hacerme responsable de mi trabajo si no sé explicarlo, ni a los demás ni a mí mismo? ¿Cómo voy a garantizar el buen funcionamiento de un producto de ingeniería que ha sido producido mediante un proceso de "caja negra", es decir, que es inexplicable (o, al menos, inexplicado hasta este momento)?

Palabras clave: inteligencia artificial, aprendizaje automático, explicabilidad, racionalidad, responsabilidad.

Abstract:

The author intends to answer the following question: what kind of responsibility can an engineer assume when using artificial intelligence tools to develop his or her work? Obviously, engineers have always used all kinds of tools: conceptual, mathematical, mechanical. So, what do artificial intelligence techniques add that might pose an additional problem? Why is this a new problem, requiring specific reflection? To put it very briefly, the problem would be this: how am I to take responsibility for my work if I do not know how to explain it, either to others or to myself? How am I to ensure the proper functioning of an engineering product that has been produced through a "black box" process, i.e. one that is inexplicable (or, at least, unexplained up to this point)?

Keywords: artificial intelligence, machine learning, explainability, rationality, responsibility.

http://www.fgbueno.es/bas
basilisco@fgbueno.es

ISSN 0210-0088 (vegetal) - ISSN 2531-2944 (digital)
Depósito Legal: O-343-78

Inteligencia artificial: explicabilidad, racionalidad y responsabilidad profesional del ingeniero

Gonzalo Génova Fuster
Universidad Carlos III de Madrid

El problema

En esta comunicación pretendo responder a la siguiente pregunta: ¿qué tipo de responsabilidad puede asumir un ingeniero que utiliza herramientas de inteligencia artificial para desarrollar su trabajo?

Obviamente, los ingenieros siempre han utilizado todo tipo de herramientas: conceptuales, matemáticas, mecánicas. Entonces, ¿qué añaden las técnicas de inteligencia artificial que pueda suponer un problema adicional? ¿Por qué esto es un problema nuevo, que requiere una reflexión específica?

Por decirlo muy brevemente, el problema sería este: ¿cómo voy a *hacerme responsable* de mi trabajo si no sé explicarlo, ni a los demás ni a mí mismo? ¿Cómo voy a *garantizar* el buen funcionamiento de un producto de ingeniería que ha sido producido mediante un proceso de "caja negra", es decir, que es inexplicable (o, al menos, inexplicado hasta este momento)?

Dos vertientes de la inteligencia artificial

El término "inteligencia artificial" agrupa en realidad una pluralidad de técnicas muy diversas, desde la aplicación de *reglas de razonamiento abstractas*, hasta el mero *descubrimiento de patrones* en grandes cantidades de datos [8]. El primer grupo de técnicas queda ilustrado por los Sistemas Expertos, en los que un "motor de inferencia" programado en un ordenador es capaz de inferir nuevas proposiciones a partir de un conjunto dado de proposiciones, que ya han sido asumidas como verdaderas. Por ejemplo, a partir de la premisa de que "tales y tales síntomas son indicativos de tal enfermedad", y la premisa de que "tal paciente presenta tales síntomas", se puede inferir (incluso especificando un grado de probabilidad) que tal paciente padece tal enfermedad. El segundo grupo, que genéricamente se puede denominar Aprendizaje Automático (*machine learning*), incluye técnicas tan variadas como visión artificial (reconocimiento de señales de tráfico, de

Gonzalo Génova Fuster. *Inteligencia artificial: explicabilidad, racionalidad y responsabilidad profesional del ingeniero*
El Basilisco, nº 59 (2023), págs. 44-48. ISSN 0210-0088 (vegetal) ISSN 2531-2944 (digital)

rostros humanos, etc.), predicción del comportamiento de consumidores a partir de su historial de búsquedas (y, en general, predicción de todo tipo de tendencias), modelos predictivos de lenguaje (tan de moda hoy día para simular conversaciones), etc.

El primer grupo se caracteriza, en cierto modo, porque las reglas de razonamiento están programadas *a priori*, mientras que en el segundo grupo el funcionamiento resultante no ha sido programado, sino que es "aprendido" *a posteriori*, a partir de patrones observados en los datos. En la práctica, la división no es tan tajante en los sistemas reales, sino que se usa una combinación de técnicas que los sitúa en algún lugar dentro de un amplio espectro entre esos dos polos. Por ejemplo, un traductor automático puede usar reglas gramaticales (programadas a priori) junto con estadísticas de traducción para determinadas expresiones (aprendidas a posteriori).

Explicabilidad y racionalidad

La fiabilidad del primer grupo de técnicas se basa en la *autoridad de los "expertos"*, que certifican que las reglas de razonamiento son correctas (naturalmente, no una autoridad absoluta, sino sometida al juicio de cualquiera que también sepa del tema). En cambio, en el segundo caso, la fiabilidad se basa en una especie de "hasta ahora siempre ha sido así, y además podemos hacer nuevas predicciones y verificar que son acertadas". Creo que puede reconocerse aquí el clásico *problema de la inducción*, o de la relación entre *correlación y causalidad*. La correlación es muy fuerte, muy sólidamente verificada, pero no somos capaces de dar una *explicación racional de la validez de las reglas* que han sido extraídas a partir de los datos. Valen, funcionan, pero no sabemos por qué (o sea, no podemos explicarlas en términos de reglas a priori).[1]

Que un sistema esté basado en reglas a priori no significa que su funcionamiento sea fácil de explicar. Si el número de reglas es grande, y la relación entre ellas es compleja, entonces el funcionamiento global del sistema puede ser extraordinariamente difícil de comprender.[2] No me refiero solamente a sistemas de procesamiento de información: lo dicho vale igualmente para grandes sistemas mecánicos, como un buque de pasajeros (el *Titanic*). No es fácil que una sola mente humana sea capaz de entender en todos sus detalles la complejidad de una obra de ingeniería de estas características.[3] Pero la cuestión clave es que la pregunta, *¿por qué este tornillo está aquí?*, se puede responder.[4] Si yo mismo no sé la respuesta, sé al menos que hay alguien que sí la sabe; es más, si aun sabiendo la respuesta no soy capaz de comprenderla completamente, sé que hay alguien que sí la comprende. Hay respuesta porque todo el proyecto de ingeniería responde a un *diseño racional*, es decir, un diseño guiado por objetivos globales, y objetivos parciales en cada una de sus partes. Análogamente, un sistema software que consista en millones de líneas de código de programa es probablemente inaccesible a una sola mente humana; pero sé que la pregunta, *¿por qué esta línea de código?*, tendrá respuesta. En definitiva, un sistema cuyo diseño esté basado en reglas a priori es explicable, si bien no necesariamente explicable por mí mismo o por una sola persona cualquiera.

Por el contrario, en los sistemas de aprendizaje automático las piezas van encajando, por así decir, evolutivamente, sin un orden preestablecido, de forma que no se puede preguntar en concreto por qué esta pieza y no otra está aquí. El ejemplo más conocido es el de las *redes neuronales*, que consisten en pequeñas unidades de cálculo (neuronas) que combinan una serie de datos de entrada mediante operaciones matemáticas muy sencillas para producir un dato de salida, que a su vez se puede usar como dato de entrada en una o varias neuronas en capas sucesivas, según el modelo matemático de McCulloch y Pitts [6]. Esas operaciones matemáticas consisten, por ejemplo, en multiplicar el dato de entrada por un factor, que viene a ser un parámetro de la neurona. De modo que el funcionamiento de la red de neuronas en su conjunto depende del valor de esos parámetros, y el proceso de "aprendizaje" (también denominado "entrenamiento") consiste precisamente en encontrar los parámetros que logran ajustar la red para que sea capaz de reconocer patrones, regularidades, en los datos de entrenamiento.[5] Las redes de neuronas que se utilizan actualmente con éxito en diversas aplicaciones constan de millones, o miles

(1) Como puede observarse, estoy dando mayor valor a la racionalidad *a priori* frente a la racionalidad *a posteriori*. Esto no implica que la segunda no tenga ningún valor, o que no sea racionalidad de ninguna manera.

(2) De hecho, incluso el funcionamiento de un sistema con muy pocas reglas puede ser muy difícil de explicar. Por ejemplo, las reglas del ajedrez son pocas y bastante sencillas, tomadas una a una. No obstante, la complejidad de juego que puede generarse es inmensa, de modo que no se ha logrado demostrar teóricamente quién sería el ganador en caso de juego perfecto por ambas partes, aunque las estadísticas (tanto en partidas humano-humano como en partidas máquina-máquina) confirman la intuición de que las blancas tienen cierta ventaja. Véase https://www.chessgames.com/chessstats.html, https://es.wikipedia.org/wiki/Ventaja_de_salida_en_ajedrez.

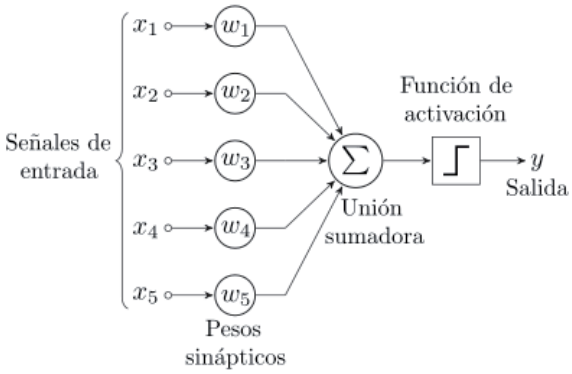
(3) Debo este ejemplo a mi gran amigo Ignacio Quintanilla Navarro, recientemente fallecido.

(4) Obviamente, una respuesta tal como "este tornillo está aquí porque lo puso Francisco, el mecánico" no sería aceptable como respuesta última. Ya que, ¿por qué Francisco puso ahí el tornillo? Es decir, el ingeniero no pregunta por la Causa de un elemento del sistema, sino por su Razón. Aunque, bien mirado, esa respuesta "histórica" también sería aceptable si la pregunta se entiende en un sentido diferente: cuando ya no pregunto por el diseño, para entender su razón de ser, sino por algo que parece un error en la implementación del diseño. Quiero saber, en definitiva, la causa del error.

(5) El patrón reconocido es el resultado de la red de neuronas, expresado matemáticamente. Por ejemplo, las coordenadas del rectángulo que encierra un rostro reconocido en la cámara de un teléfono móvil, la clasificación de un comportamiento como impulsivo o sosegado, la siguiente palabra más probable que escribirá el usuario en la caja de texto, etc.

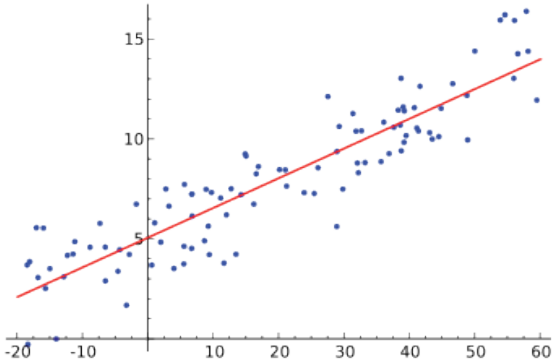
Gonzalo Génova Fuster. *Inteligencia artificial: explicabilidad, racionalidad y responsabilidad profesional del ingeniero*
El Basilisco, nº 59 (2023), págs. 44-48. ISSN 0210-0088 (vegetal) ISSN 2531-2944 (digital)

de millones, de neuronas con sus respectivos parámetros. Como es fácilmente entendible, para llegar a encontrar ese enorme número de parámetros que hacen que la red de neuronas funcione se requiere un proceso conceptualmente complicado y computacionalmente muy costoso.



Modelo matemático de una neurona artificial

No obstante, la técnica puede ilustrarse con el método matemático de *regresión lineal*, mediante el cual una nube de puntos que representa la relación entre dos variables (tales como altura y peso de una persona) se aproxima mediante una línea recta, cuya ecuación algebraica se define, como es bien sabido, mediante dos números.[6] La justificación de esos dos números es… que definen una línea recta que *se ajusta bien* a la nube de puntos. No hay más. No hay otra racionalidad que la racionalidad del ajuste estadístico de la solución al patrón de los datos de entrenamiento. En definitiva, lo que se produce con el aprendizaje automático es un sistema que imita el conjunto de datos con los que ha sido entrenado. *Lo que se garantiza es la semejanza, la imitación.*[7] No se produce ni garantiza ningún otro tipo de racionalidad, de diseño intencional.



Ejemplo de regresión lineal con una variable dependiente y una variable independiente

(6) Es decir, la recta de regresión es una función matemática que puede ser parametrizada para que se ajuste a un conjunto cualquiera de datos. En este sentido, la red neuronal es algo parecido a la recta de regresión –una función matemática parametrizable–, pero millones de veces más complicado.

(7) Pienso que es muy significativo que el mismo Alan Turing comience su famoso artículo de 1950 sobre la inteligencia artificial [9] planteando el problema de la máquina pensante como un juego de imitación. *The Imitation Game* es el título del primer apartado de ese artículo, y es también el título del biopic de 2014 (en España fue titulado *Descifrando enigma*).

Racionalidad imitativa

Las carencias de esta *racionalidad imitativa* se ponen de manifiesto de modo particular en las discusiones sobre la ética de la inteligencia artificial [4]. En este campo particular, la explicabilidad preocupa mucho a los investigadores, y también ya –por fin– a los gobiernos [2]. El resultado final del aprendizaje automático es una fórmula, un algoritmo, para reconocer un patrón en el que encajan rostros, tendencias, series de preguntas y respuestas, pero *no se puede explicar por qué la fórmula funciona*; no hay otra justificación para ella fuera de que tenga un porcentaje de éxito muy elevado, su efectividad en el ajuste con los datos de entrenamiento y con la predicción de nuevos sucesos. Supongamos ahora que esta semejanza con el patrón aprendido se usa para tomar una decisión. Pues bien, cuando esta decisión tiene una fuerte carga ética, el hecho de que no se pueda justificar *razonadamente* es un problema serio, muy serio.

Y esto ocurre no solo en sistemas totalmente automatizados, como los muy mediáticos y ya no tan futuristas *vehículos autónomos*, que supuestamente deben decidir por sí mismos a quién deben atropellar, basados en la imitación del criterio "moral" de millones de personas encuestadas [1]. El problema se presenta también en sistemas que se limitan a asesorar a un agente humano en sus decisiones, tal como un sistema que recomienda contratar a una persona, conceder un préstamo bancario, o incluso otorgar la libertad condicional a un preso. En este caso la decisión última la toma el agente humano, pero aun así queda el problema, entre otros muchos, de que no basta con dar una recomendación, sino que esta debe ser *razonada*. El funcionario de Hacienda puede recibir la sugerencia de que un contribuyente está defraudando en su declaración de impuestos; el profesor puede recibir la sugerencia de que el ensayo del alumno no es original. Pero, en ambos casos, no basta con la sugerencia para multar o para suspender: hay que ser capaz de demostrar el fraude con pruebas fehacientes. *No es aceptable tomar una decisión moral con base en el resultado de una caja negra que ha proporcionado un número no verificable.*

Así pues, a menudo ocurre que los sistemas de inteligencia artificial son tan complicados que es muy difícil saber por qué han llegado a una determinada conclusión. Si el profesional asesorado es sensato, rechazará las recomendaciones del sistema que no estén bien argumentadas, que estén aquejadas de falta de transparencia o "explicabilidad". Desafortunadamente, no es desdeñable tampoco el riesgo de proliferación de profesionales acomodaticios, que acepten una decisión anónima sin cuestionarla, y por tanto sin verdadera capacidad de hacerla propia.

Gonzalo Génova Fuster. *Inteligencia artificial: explicabilidad, racionalidad y responsabilidad profesional del ingeniero*
El Basilisco, nº 59 (2023), págs. 44-48. ISSN 0210-0088 (vegetal) ISSN 2531-2944 (digital)

Como hemos visto, la dificultad para explicar las decisiones artificiales puede presentarse tanto si el sistema está basado en reglas programadas a priori, a partir de un diseño racional, como si está basado en reglas aprendidas a posteriori, a partir de la imitación de un patrón observado en los datos. Pero la diferencia en las raíces de la inexplicabilidad en un caso y otro es crucial. En el primer caso la falta de explicabilidad es meramente práctica y relativa: como decía antes, se debe a la complejidad del sistema, que lo hace difícil, pero no imposible, de entender y explicar para según qué personas; porque, de hecho, es un sistema *diseñado racionalmente*, y por tanto en sí mismo explicable. Por el contrario, en el segundo caso estamos ante una falta de explicabilidad esencial, porque *no hay diseño racional*, tan solo hay imitación de un comportamiento observado.

El mero reconocimiento de un patrón en un conjunto de fenómenos no permite asegurar que haya una intención de diseño que dé origen a esas regularidades.[8] Y si no hay o no se conoce la intención de diseño, o sea, la Razón, entonces la imitación de ese patrón no puede servir como base para un comportamiento responsable. O, al menos, resulta muy problemática como base para asumir responsabilidades profesionales.

Supongamos que uso un reconocedor de patrones para emular la forma en la que un ingeniero software evalúa la calidad de los requisitos de un proyecto (si se entienden, si son coherentes, si son completos, etc.), a partir de determinadas magnitudes medibles en el texto de los requisitos (longitud de las frases, terminología empleada, relaciones de unos requisitos con otros, etc.). Incluso si la herramienta llega a tener un elevado porcentaje de concordancia con los datos de entrenamiento, lo que no puede hacer es explicar verdaderamente *por qué el ingeniero experto clasificó los requisitos de muestra en buenos y malos*; o sea, no puede explicar sus "razones" para hacerlo así. Entonces, si yo uso ese emulador de experto en mi propio proyecto de desarrollo de software, podré estar razonablemente seguro de que soy capaz de identificar requisitos mal escritos, pero en realidad no soy capaz de justificarlo. Y esto es el *quid*: ¿qué clase de ingeniero soy si no puedo justificar racionalmente mi trabajo, qué responsabilidad puedo asumir?

Maticemos un poco. Se puede admitir que la racionalidad imitativa es una cierta forma de racionalidad. Si yo, ingeniero, actúo conforme a lo que es comúnmente aceptado en mi profesión (es decir, imito a otros), aunque no entienda completamente esa forma de proceder, tal vez no se puede decir que sea un irresponsable. Ahora bien, la cuestión es que, si yo hago lo que he aprendido en un libro, aunque no lo entienda todo, puedo preguntar al que escribió el libro. Pero si hago lo que dice una máquina que ha llegado a un resultado por un proceso de ajuste paramétrico... no hay nadie que pueda dar razón de la respuesta, *no puedo preguntar a nadie*, porque la respuesta no se ha generado de forma racional, motivada. O, más bien, su única racionalidad es el ajuste estadístico con una gran muestra de datos. Si se cae el puente porque fue diseñado como dijo la máquina, pero nadie entiende verdaderamente ese diseño, ¿quién puede hacerse responsable?

Responsabilidad profesional

El problema de la racionalidad imitativa salta a la palestra, como decía, específicamente al considerar la explicabilidad de las decisiones con carga moral que queramos delegar en un sistema de inteligencia artificial. Ahora bien, en realidad el problema se presenta en cualquier sistema cuyo funcionamiento tenga consecuencias éticas. ¿Y qué sistema no las tiene? Pongamos por caso un ascensor cuyas puertas automáticas renuncian a cerrarse cuando hay un objeto que lo impide (típicamente, una persona o alguna de sus pertenencias). Con esto se está incumpliendo la misión del ascensor, que es trasladar a sus pasajeros a la mayor brevedad al piso elegido. ¿Por qué? Porque hay un requisito "más importante", que es no lastimar a esos pasajeros. ¿Y por qué es más importante? Es obvio, pero no por eso deja de ser verdad, que es un *requisito ético implícito* en el diseño del ascensor.

Toda actividad profesional tiene implicaciones éticas, no solamente aquellas que se enfrentan a "difíciles dilemas éticos", tal como a menudo se presentan al gran público. Y así llegamos a las preguntas que motivan esta reflexión: ¿cómo voy a hacerme responsable de mi trabajo si no sé explicarlo, ni a los demás ni a mí mismo? ¿Cómo voy a garantizar el buen funcionamiento de un producto de ingeniería que ha sido producido mediante un proceso de "caja negra", es decir, que es inexplicable (o, al menos, inexplicado hasta este momento)?

La inteligencia artificial generativa se ha puesto de moda con sistemas como ChatGPT y LaMDA (sistemas conversacionales), o Dall·E y Stable Diffusion (generación de imágenes a partir de una descripción textual), o incluso AlphaCode (un sistema de generación automática de programas de ordenador). Cuando estos sistemas se usan como mero entretenimiento, tal vez las cuestiones éticas sean menos relevantes. Pero, en el momento en

(8) En biología, los partidarios del *Intelligent Design* sostienen justamente la tesis contraria. No obstante, el método científico-experimental moderno en sentido estricto sigue la conocida tesis baconiana de que de que la ciencia debe limitarse a estudiar los fenómenos naturales y sus regularidades, sin tratar de descubrir ninguna finalidad o diseño en ellos: *nam causarum finalium inquisitio sterilis est, et, tanquam virgo Deo consecrata, nihil parit* ("la investigación de las causas finales es una cosa estéril, no parirá nada, igual que una virgen consagrada a Dios"), Francis Bacon, *De Augmentis Scientiarum*, III, 5 (1623). Otra cosa es que pongamos en ejercicio una racionalidad que vaya más allá de lo puramente empírico y verificable.

Gonzalo Génova Fuster. *Inteligencia artificial: explicabilidad, racionalidad y responsabilidad profesional del ingeniero*
El Basilisco, nº 59 (2023), págs. 44-48. ISSN 0210-0088 (vegetal) ISSN 2531-2944 (digital)

que la inteligencia artificial se usa como herramienta en una actividad profesional, la cuestión es ya insoslayable. Tomemos el caso de la producción de software. ¿Qué empresa certificará un producto software del que no puede dar una completa explicación? ¿Qué empresa se comprometerá con sus clientes en la calidad de un software que es una pura caja negra? Es obvio que no basta decir "la máquina dijo que...". Y justo por eso productos como ChatGPT son tan peligrosos, al menos mientras nadie asuma verdadera y plenamente la responsabilidad por las respuestas que da. La industria seria es consciente del problema, y se trabaja mucho en encontrar formas de incrementar la interpretabilidad y confiabilidad en las respuestas de la inteligencia artificial [7].

Reflexión final: la racionalidad en la ciencia y en la tecnología

Entender y definir qué sea la racionalidad es una tarea que excede con mucho las pretensiones de esta comunicación. Pero sí espero que las reflexiones precedentes sirvan para ilustrar una importante diferencia entre la racionalidad en la ciencia y en la tecnología (es decir, con el uso teórico y el uso productivo de la razón). Por mucho que la revolución científica moderna haya sido simultáneamente una revolución tecnológica, hay que señalar que ciencia y tecnología, en su mutua dependencia, no son lo mismo.

Aun a riesgo de simplificar excesivamente, la racionalidad científica consiste básicamente en la regularización de los fenómenos de la naturaleza en leyes que sirven para hacer predicciones. Que estas leyes sean expresión matemática de verdaderas causas físico-mecánicas, o simplemente correlaciones entre fenómenos observables, es una cuestión en la que no voy a entrar. En todo caso, hay un cierto parentesco con los patrones que somos capaces de encontrar en los datos mediante las técnicas de inteligencia artificial: observación-regularización. Por el contrario, la racionalidad tecnológica es inseparable de la consideración del diseño y finalidad de los artefactos [5]. La racionalidad tecnológica no se conforma con establecer leyes universales, regularidades; las aprovecha, sin duda, pero no es reducible a ellas.

Podemos pedir responsabilidad a los agentes que obran por "razones" (es decir, intenciones racionales), pero no a aquellas entidades que obran meramente por "causas" (interacciones físico-mecánicas). Por este motivo hay también dos géneros diferentes de explicabilidad: explicar para predecir/controlar (explicación de sistemas físicos o cuasi-físicos, tales como grandes masas sociales), y explicar para pedir responsabilidades (explicación legal o moral). Estos dos géneros se corresponden a su vez con la explicación por Causas y la explicación por Razones.[9]

Para terminar, quiero simplemente recapitular la tesis central que he defendido en este artículo. No es aceptable tomar una decisión moral con base en el resultado de una caja negra que ha proporcionado una respuesta no verificable ni explicable. Y, puesto que todo producto de un proceso de ingeniería tiene consecuencias éticas, no es aceptable hacer ingeniería de caja negra. Esto es el *quid*: ¿qué clase de ingeniero soy si no puedo justificar racionalmente mi trabajo, qué responsabilidad puedo asumir? Un ingeniero no puede hacerse responsable de un trabajo que no sabe explicar, ni a los demás ni a sí mismo. Por lo tanto, para desarrollar una ingeniería éticamente responsable no basta con herramientas de IA que respondan a una racionalidad *a posteriori* (ajuste a los datos de entrenamiento); hacen falta herramientas que respondan a una racionalidad *a priori* (conformidad con un diseño racional, intencionado).

Agradecimientos

Agradezco las sugerencias de Elizabeth Scherschener, Artem Urlapov y Jesús Poza sobre un borrador previo de esta comunicación.

Referencias

[1] Awad, E., Dsouza, S., Kim, R., Schulz, J., Henrich, J., Shariff, A., Bonnefon, J.-F., & Rahwan, I. (2018). The moral machine experiment. *Nature* 563:59–64.

[2] Comisión Europea, Grupo de expertos de alto nivel sobre inteligencia artificial (2019). Directrices éticas para una IA fiable (Ethics guidelines for trustworthy AI). https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

[3] García Norro, J.J. (2012). ¿Es natural la inteligencia? En M. Oriol (ed.), *Inteligencia y filosofía*. Madrid: Marova, pp. 151-169.

[4] Génova, G., González, M. R., Moreno, V. (2022). A lesson from AI: Ethics is not an imitation game. *IEEE Technology and Society Magazine* 41(1):75–81.

[5] Génova, G., Quintanilla Navarro, I. (2018). Discovering the principle of finality in computational machines. *Foundations of Science* 23(4):779–794.

[6] McCulloch, W.S., Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *Bulletin of Mathematical Biophysics* 5(4):115–133.

[7] Molnar, C. (2019). *Interpretable Machine Learning: A Guide For Making Black Box Models Explainable*. Morrisville, NC: Lulu.

[8] Russell, S., Norvig, P. (2009). *Artificial Intelligence: A Modern Approach*. Englewood Cliffs, NJ: Prentice Hall, 3.ª ed.

[9] Turing, A.M. (1950). Computing machinery and intelligence. *Mind* 59:433-460.

Recibido: 02-09-23

Aceptado: 20-10-23

(9) Sobre la distinción entre causa como explicación físico-mecánica y razón como justificación lógica, véase también García Norro, ¿Es natural la inteligencia? [3].