Geppetto moldea a Pinocchio: la metáfora del semáforo aplicada al análisis de la IA

Alberto Fernández Fernández. Universidad de Oviedo (España)
Dr. Román García Fernández. *Eikasía, Revista de Filosofía* (España)
Recibido 28/06/2025 • Aceptado 28/10/2025

Resumen

Un viejo carpintero construye un títere de madera que aspira a ser «un niño de verdad». El desarrollo tecnológico permite generar artefactos prodigiosos pero ¿son de verdad? Aplicamos nuestros tres momentos metodológicos (fenoménico, positivo y valorativo) al análisis del concepto de inteligencia artificial, y proponemos una sentencia moral: ¿tras Copérnico, Darwin y Freud, es esta la cuarta afrenta a la singularidad humana?

Palabras clave: artefactos, conciencia, contrafácticos, metáfora del semáforo, inteligencia artificial, posthumanismo, robots, singularidad humana, verdad.

Abstract

Geppetto shapes Pinocchio: the traffic light metaphor applied to the analysis of AI

An old carpenter builds a wooden puppet that aspires to be «a real boy». Technological development allows the generation of prodigious artifacts, but are they real? Truly real? We apply our three methodological moments (phenomenical, positive and evaluative) to the analysis of the concept of Artificial Intelligence, and propose a moral sentence: after Copernicus, Darwin and Freud, is this the fourth affront to human singularity?

Keywords: artifacts, consciousness, counterfactuals, traffic light metaphor, artificial intelligence, posthumanism, robots, human singularity, truth.

ISSN-e: 1885-5679
V Congreso Iberoamericano de Filosofía de la Ciencia y la Tecnología

ISSN-e: 1885-5679
V Congreso Iberoamericano de Filosofía de la Ciencia y la Tecnología

Geppetto moldea a Pinocchio: la metáfora del semáforo aplicada al análisis de la IA

Alberto Fernández Fernández. Universidad de Oviedo (España)
Dr. Román García Fernández. *Eikasía, Revista de Filosofía* (España)
Recibido 28/06/2025 • Aceptado 28/10/2025

Una cámara es una máquina, y a las máquinas
se las fuerza, se las pega, se las maltrata.
¡Lo importante es la poesía!
Orson Welles, *Ciudadano Welles*

§ 1. Pinocchio quiere ser un niño de verdad

La utilización masiva de herramientas de inteligencia artificial a partir de 2023 como Google Translate, ChatGPT o Whisk entre la mayoría de mi alumnado de bachillerato me ha forzado a plantearme, como profesor de filosofía en ejercicio, algunas preguntas sobre la esencia de esta novedosa tecnología, sobre su sentido epistémico y sobre la bondad, o no, de sus aplicaciones prácticas. Nosotros hemos tratado de sistematizar toda posible casuística del término en una contribución publicada en esta misma revista que realiza, en la medida de nuestras fuerzas, una precisa, rigurosa y sucinta cartografía del nombre que tratamos de delimitar: ¿qué es la «inteligencia artificial»? (García Fernández y Fernández Fernández, 2026).

Nos gusta comenzar nuestras presentaciones públicas «al modo platónico», proponiendo al oyente un «mito». Los alumnos del IES Pando de Oviedo (Asturias) denominan al ChatGPT Geppetto, que como bien sabrán ustedes es un personaje de ficción ideado por Carlo Collodi para su novela *Le avventure di Pinocchio* (*Las aventuras de Pinocho*) publicada en la ciudad de Florencia en 1883 —un texto que, por cierto, no es en absoluto un cuento infantil, por más que en el imaginario colectivo haya prevalecido la interpretación naif de 1940 del animador Ben Sharpsteen para la productora Walt Disney—, un mito moderno que ha sido llevado al cine en múltiples ocasiones, una de las más recientes la realizada por el cineasta Steven Spielberg en la bienintencionada aunque fallida *A.I. Artificial Intelligence* (Warner Bros, 2001).

ISSN-e: 1885-5679
V Congreso Iberoamericano de Filosofía de la Ciencia y la Tecnología

Geppetto es un «artesano» que, al igual que el *demiurgo* platónico, no crea la realidad desde la nada, sino que la ordena a partir de una masa caótica preexistente, en este caso madera y hebras, para construir una marioneta, una materia con forma humana pero carente de alma, a la que da el nombre de Pinocchio. Pero el títere aspira a tener un alma, una identidad y una personalidad propias: quiere ser «un niño de verdad». Tres inconvenientes surgen a tal respecto: el primero es que Pinocchio dispone de una presunta «conciencia» (Pepito Grillo) que no le es inherente, que le viene impuesta desde fuera, de forma azarosa, sobrevenida, cuántica —toda marioneta pende de unos hilos, que son manejados desde el exterior—; la segunda es que carece de sexo, lo que le impide desarrollarse como un ser corpóreo operatorio en sentido pleno —es incapaz de reproducirse, de perpetuarse, de procurar una nueva generación—; la tercera, y quizás la más significativa de todas, es que Pinocchio miente… y lo hace sistemáticamente, ya sea como mecanismo de defensa o como herramienta de supervivencia —como refiere Voltaire en esta misma línea: «Yo, como Don Quijote, me invento pasiones para ejercitarme», esto es, me miento a mí mismo para ser feliz.

Profundicemos el mito: un artesano, Geppetto —la tecnología de aprendizaje profundo GPT (*Generative Pre-trained Transformer*)—, ha desarrollado un producto, Pinocchio —un *chatbot* que genera textos de forma automática haciendo uso de NLP (*Natural Language Processing*)—, una novedosa forma de «inteligencia» que potencialmente podría alcanzar el estado de «conciencia». ¿Es esto posible «de verdad»? ¿Apunta la IA a una nueva realidad óntica? ¿Estamos ante una opción real o ante una quimera?

Analicemos los acontecimientos: el desarrollo técnico ha permitido generar artefactos prodigiosos desde la Antigüedad —los inventos mecánicos de Herón de Alejandría, la máquina deductiva de Ramón Llull, los aparatos bélicos de Leonardo da Vinci, la *pascalina* de Blaise Pascal y la *mathesis universalis* de Godofredo Leibniz, los autómatas de la Ilustración, la máquina de vapor de James Watt, el procedimiento algorítmico de Ada Byron, y un largo etcétera—. Este progreso tecnológico alcanza su epítome en la era postindustrial de la computación, con el desarrollo de ordenadores, redes y algoritmos que casi parecen obra de magia. La operatividad de estos nuevos artefactos de conocimiento nos obliga a una reflexión serena y precisa.

ISSN-e: 1885-5679
V Congreso Iberoamericano de Filosofía de la Ciencia y la Tecnología

§ 2. La metodología de la metáfora del semáforo

Proponemos una metodología de análisis denominada «metáfora del semáforo» que se articula en tres momentos —fenoménico, positivo y valorativo— y que trataremos de aplicar al análisis del concepto de inteligencia artificial, y sus inquietantes promesas de futuro. El nuevo método busca integrar la mayéutica socrática, el análisis/síntesis cartesiano, la geometría moral espinosista, la tríada dialéctica hegeliana y la praxis moral marxista. Será necesario para ello definir términos (conceptualizar), ordenarlos de forma sistemática (categorizar) y explicitar sus relaciones lógicas (argumentar).

Partimos de una sencilla metáfora visual para ejemplificar este método: un semáforo —que es un robot primitivo, un artefacto fácil de identificar por nuestro alumnado—. Un primer momento nos obligaría a estar detenidos, con la luz en rojo: es el momento en el que estamos a la expectativa, y aceptamos todo aquello que vemos de forma pasiva, acrítica. En un segundo momento, en cuanto la luz cambia a ámbar, nos vemos obligados ya a prestar más atención, a mirar con detenimiento a uno y otro lado, a fijarnos en las cosas de forma más activa, en otras palabras, a practicar algún tipo de análisis de la situación. Y finalmente, en un tercer momento, la luz se pone en verde y nos permite ponemos en marcha, llevar a la práctica todo lo que hemos visto y aprendido con anterioridad y ejecutar nuestra acción.

El primer momento, —fenoménico— formula el problema en términos de conocimiento de sentido común, desde los saberes mundanos disponibles, muchos de ellos saberes bárbaros, como los mitos o la religión, o civilizados acríticos, como las pseudociencias o las ideologías, a objeto de identificar las fuentes de información y discriminar unas de otras para poder evaluarlas. La mayoría de nosotros no sobrepasa este primer momento, ya que acepta la realidad tal y como la percibe a través de los medios de comunicación de masas, que contribuyen a generar un «efecto ruido» sobre el tema e inundan nuestras cabezas de estereotipos y prejuicios, lo que nos obligará a ejercitar una radical distinción entre las «apariencias» y los «fenómenos».

El segundo momento, positivo —esto es, gnoseológico— de análisis en profundidad del conflicto desde las aportaciones de los saberes civilizados, muy especialmente desde las ciencias (formales, naturales, sociales) y las tecnologías, que nos permita

investigar con rigor académico y en profundidad el objeto de estudio que nos hemos propuesto. Deberemos echar mano de las aportaciones de la matemática, la lógica, la física, la informática, la computación… pero también de la biología, la anatomía, la antropología, la psicología, la sociología, el derecho… si de verdad queremos desentrañar el sentido de los términos, operaciones y relaciones generadas. En último término, deberemos ejercitar la distinción platónica entre la «doxa» y la «episteme».

El tercer momento, valorativo —esto es, ético, práctico o propiamente filosófico—, nos permitirá valorar, desde las distintas teorías éticas disponibles, algunas posibles soluciones al conflicto para poder argumentar racionalmente y generar nuestro propio juicio o sentencia moral sobre el tema «al modo socrático», mediante una «definición». Será desde la dialéctica ética-política-moral desde la que podremos desentrañar el sentido genérico, esencial, del concepto objeto de estudio. La diferencia entre éticas «intencionales» y éticas «consecuencialistas» resultará crucial, y nos obligará a un posicionamiento «materialista formalista» para proponer soluciones al problema, que deberán articularse de alguna manera en una práctica política efectiva, sólida y viable.

Para una aproximación más detallada a este método de trabajo, remitimos al artículo publicado en la revista del Instituto de Estudios para la Paz y la Cooperación bajo el título «La metáfora del semáforo» (Fernández Fernández, 2010). En ese mismo artículo especificamos además una clasificación de los conflictos morales en tres grandes rúbricas: problemas personales e interpersonales; problemas sociales, políticos y económicos; y problemas tecnológicos o de relación con el medio ambiente. En esta tercera categoría cabría situar el conflicto ético-político-moral planteado por el reciente surgimiento, desarrollo y aplicación de la inteligencia artificial, problema que nosotros entendemos como no resuelto —y que tal vez resulte irresoluble—, y sobre el que trataremos de arrojar alguna luz en las siguientes líneas.

§ 3. La percepción fenoménica de la inteligencia artificial

¿Cómo percibimos la inteligencia artificial la mayoría de las personas? Apliquemos el método, comenzando por el momento fenoménico, que es el modo en que los individuos y las culturas perciben y simbolizan la tecnología antes de comprenderla.

Primeramente, a través de mitos y relatos culturales como el mito griego de Prometeo —que roba el fuego de los dioses y otorga a los hombres el poder de la técnica— o el mito bíblico de la expulsión del paraíso —en el que el trabajo se nos impone como un castigo divino del que el ser humano intenta liberarse construyendo máquinas que le suplanten—. En segundo lugar estarían las tradiciones, fundamentalmente la idea grecolatina de «excelencia» (*areté*) y la idea ilustrada de «progreso» (*progrès*), que al igual que los mitos precedentes remiten en última instancia a la idea de «mejoramiento de lo humano», pero que por desgracia nos fuerzan a desarrollar una serie de estereotipos y perjuicios que podemos resumir en lo que genéricamente llamamos la disputa sobre la «bondad de la tecnología» y la «maldad de la tecnología» —entre las vacunas que salvan vida y las bombas que las arrebatan, por decirlo sucintamente.

Esta dicotomía —tecnofilia frente a tecnofobia— constituye una disputa casi metafísica entre optimismo y catastrofismo que atribuye a la IA virtudes o peligros desmesurados, lo que resulta del todo estéril y no ayuda a solucionar el problema. El resultado de esa oscilación, estimulada por los medios de comunicación de masas con su efervescente «efecto ruido», es un estado de «apariencias» que nos impide comprender la realidad. En el caso de la IA, este estado apariencial es especialmente dramático por obra y gracia de los relatos de ciencia ficción, tanto literarios —*Un mundo feliz*, *Solaris*, *¿Sueñan los androides con ovejas eléctricas?*, *Yo, Robot* o *Ghost in the Shell*, entre otros— como fílmicos —*Metrópolis*, *2001: una odisea del espacio*, *Terminator*, *Matrix*, *Ex Machina*, y un largo etcétera—, que nos proveen de apasionantes y desconcertantes «contrafácticos» —como MARIA, HAL 9000, NEXUS 6, T-800 o AVA—. Y aunque estas ficciones pueden procurarnos interesantes elementos para la reflexión, lo cierto es que no hacen otra cosa sino ensombrecer nuestra percepción de la realidad.

Tenemos que superar este momento apariencial y adentrarnos en los «fenómenos» realmente existentes. La primera referencia obligada sería Enigma (1918), la máquina de rotores diseñada para cifrar y descifrar mensajes y utilizada por la armada alemana a partir de 1923, o Colossus (1944), la computadora electrónica que desencriptó los mensajes de Lorenz (1942), la máquina utilizada por los nazis en la Segunda Guerra Mundial. Avanzando un poco más en el tiempo nos encontramos con ENIAC (1955),

ISSN-e: 1885-5679
V Congreso Iberoamericano de Filosofía de la Ciencia y la Tecnología

acrónimo de *Electronic Numerical Integrator And Computer* (*Computador e Integrador Numérico Electrónico*), que probablemente pueda ser considerado con rigor el primer computador. Más tarde llegaría ELIZA (1964), el primer programa informático, y poco después Apple I (1976), el primer ordenador personal. Muchos de nosotros recordamos a Deep Blue (1996), la supercomputadora de IBM que fue capaz de derrotar al maestro de ajedrez Gary Kasparov, y también a ASIMO (2000), acrónimo de *Advanced Step in Innovative Mobility*, el «robot humanoide» (*androide*) desarrollado por la compañía Honda. Ya en nuestro siglo, cabría reseñar la aparición del dispositivo multifunción iPhone (2007), o del asistente personal ALEXA (2014). Estas serían solo algunas muestras de un enorme listado de artefactos tecnológicos que, para bien o para mal, han modificado nuestra forma de percibir y comprender la realidad, de comunicarnos entre nosotros y actuar en el mundo.

Si se nos permite una nota biográfica, el primer artefacto de este tipo con el que tuvimos contacto fue el ordenador de 8 bits ZX Spectrum, lanzado al mercado por la compañía británica Sinclair en 1982, y con el que fuimos capaces de diseñar un «cronómetro» gracias a los rudimentarios conocimientos de lenguaje de programación BASIC. Este detalle, que puede parecer frívolo, incluso vanidoso, no lo es en absoluto. Ni ustedes ni nosotros conocimos a Enigma o Colossus, a ENIAK o ELIZA. Esta pérdida de memoria técnica, unida a la superficialidad con la que la sociedad se aproxima a la historia de sus herramientas, constituye una nueva forma de mito: la creencia en una tecnología sin genealogía, sin contexto, sin responsabilidad histórica. Aunque hemos superado las apariencias y hemos avanzado hacia los fenómenos, estamos encerrados en un solipsismo que nos impide deshacernos de nuestras vivencias personales, lo que nos niega la posibilidad de comprender plenamente el problema que acometemos y nos imposibilita su resolución. Es hora de dar un paso adelante y adentrarnos en el momento positivo.

§ 4. El análisis positivo de la inteligencia artificial

¿Cómo entendemos la inteligencia artificial desde las disciplinas científicas actuales? Desarrollemos el momento positivo, que es el modo en que los individuos y

ISSN-e: 1885-5679
V Congreso Iberoamericano de Filosofía de la Ciencia y la Tecnología

las culturas comprenden la tecnología una vez analizada con rigor técnico y académico.

Son muchos los autores que han defendido que la inteligencia artificial representa la «cuarta afrenta a la singularidad humana», el cuarto desprecio a nuestra inútil vanidad. Se suele aludir con esta expresión a las grandes rupturas del pensamiento occidental que obligaron al ser humano a revisar su lugar en el cosmos y a repensar su propia condición. La primera afrenta nos la proporcionó Copérnico, al desplazar a la Tierra del centro del universo; la segunda la articuló Darwin, al situar al ser humano dentro del reino animal, como un semoviente más; la tercera la infligió Freud, al mostrar que la conciencia no domina por completo nuestra mente y nuestra conducta; y la cuarta, parece ser, sería la amenaza que nos impone la inteligencia artificial, al demostrar que la inteligencia no es un privilegio exclusivo del ser humano, que no somos las únicas entidades pensantes. Este planteamiento, más que un lamento, es una constatación crítica: el progreso técnico ha descentrado la identidad humana, nos ha desnortado.

Nada nuevo bajo el sol, por otro lado. También el descubrimiento de América supuso un reto filosófico a la definición de nuestra esencia como especie. La constatación empírica, fenoménica, de que existían «modelos humanos no clásicos» obligó a relativizar y finalmente a disolver la concepción del hombre del humanismo, lo que posibilito la construcción de una «nueva conciencia del hombre» que trataba de especificar sus notas esenciales y establecer su invariable naturaleza con los primeros tratados *De Homine*, germen de la futura antropología filosófica. La reciente polémica sobre el transhumanismo y el posthumanismo nos vuelve a posicionar en la casilla de salida, y nos obliga a preguntarnos de nuevo si es posible una inteligencia no conformada a partir de carne y hueso, sino de ceros y unos, no sujeta al carbono, sino al silicio.

El tradicional cuento «¡Que viene el lobo!», aplicado al hecho que nos ocupa, alude al temor colectivo ante el poder creciente de la tecnología —una aprensión que la literatura y los medios alimentan mediante una retórica a medio camino entre la fascinación y la alarma—. Las palabras «inteligencia artificial» aparecen en letras bien grandes en las portadas de cientos de libros de divulgación, porque es un tema de moda que impone su presencia, y que nos posiciona en una dicotomía ya planteada

ISSN-e: 1885-5679
V Congreso Iberoamericano de Filosofía de la Ciencia y la Tecnología

por Umberto Eco en su obra *Apocalípticos e integrados*: demonizar la IA o abrazarla, resistir insurgentes o asumir lo inevitable. Sin embargo, ese discurso mediático, dominado por el sensacionalismo, mantiene a la sociedad en un nivel de opinión común, sin profundidad filosófica ni análisis crítico. Nos encontramos en el ámbito de la «doxa», del relato literario, de la ficción romántica, de la verborrea.

El análisis de documentación científica académica precisa y el diagnostico de expertos reconocidos puede ayudarnos a superar este momento dóxico y profundizar en la «episteme», tanto desde revistas generalistas como las clásicas *Science* o *Nature* (por citar solo las dos más conocidas) como desde revistas más especializadas como *Progress in AI* o *IA Communications*, además de ensayos más amplios como los de Clifford A. Pickover (2021), Cade Metz (2022) o Jaume Miralles Solé (2020), por poner solo algunos ejemplos bien conocidos, que supondrían una base firme para establecer un análisis riguroso sobre el estado de la cuestión.

Los saberes positivos nos aportarían nuevas perspectivas: la biología nos permitiría indagar en el concepto de «vida orgánica», y en el hecho de que esta pudiera haber surgido a partir de materia inorgánica —una posibilidad que se conoce técnicamente como «abiogénesis»—; la psicología, que incorporaría al debate conceptos clave como «conciencia», «inteligencia» y «creatividad»; la antropología, que propone un estudio de la idea de «evolución cultural» —siguiendo la estela de autores como Marvin Harris—, o la sociología, que articula conceptos como «conflicto y cambio social» —en la línea de Ralf Dahrendorf—; las neurociencias, que ofrecen interesantes aportaciones como las de Roger Sperry o David Hubel sobre neurología y simulación de conciencia; e inevitablemente las matemáticas, esenciales para el tema que nos ocupa, que definen conceptos trocales como álgebra lineal, cálculo, probabilidad y estadística, amén de las tecnologías derivadas de todas ellas, como la informática, la computación y la robótica —que en gran medida se han segregado de su *alma mater* y han evolucionado como disciplinas autónomas.

A ello habremos de añadir un cuidadoso escrutinio de la legislación sobre IA vigente en la actualidad —nosotros nos centraremos en el Estado español y en la Unión Europea— para determinar qué se entiende realmente por inteligencia artificial en sentido legal, y hasta qué punto los distintos Estados implementan estas legislaciones a nivel nacional o supranacional. El discurso se desplaza entonces hacia el plano

ISSN-e: 1885-5679
V Congreso Iberoamericano de Filosofía de la Ciencia y la Tecnología

normativo. España, como el resto de países integrantes de la Unión Europea, asume el marco legal especificado por la misma en su *Libro blanco de la inteligencia artificial* (2020), que se vehicula en los siguientes supuestos: «supervisión, control y evaluación como principios rectores orientados hacia la excelencia y la confianza», dos conceptos que expresan el deseo de compatibilizar el desarrollo tecnológico con la seguridad ética; y más adelante en el *Acta sobre IA* para la UE (2021) enmendada y acordada en 2023 y ratificada por la Comisión Europea en 2024 —aunque se espera que no entre en vigor hasta 2026.

El Estado español, por su parte, ha redactado el Real Decreto 729/2023, de 22 de agosto, sobre el Estatuto de la Agencia Española de Supervisión de Inteligencia Artificial; así como el Real Decreto 817/2023, de 8 de noviembre, que establece un entorno controlado que implanta normas armonizadas en materia de inteligencia artificial. Ambos textos están diseñados en base a los principios europeos de confianza y excelencia antes aludidos, lo que nos plantea un problema, pues estas dos palabras encierran también una paradoja: la confianza no se decreta, se construye, y la excelencia no es un valor moral sino técnico. En último término, el lenguaje jurídico revela las tensiones entre regulación, responsabilidad y autonomía de la técnica. El debate sobre la IA, por tanto, no debería limitarse a la política o a la ingeniería, sino que debería abordarse también desde la filosofía práctica.

Resumiendo lo dicho hasta ahora, tendríamos en primer lugar una inteligencia artificial originaria o simbólica, de carácter deductivo y lógico, basada en reglas explícitas; y en segundo lugar una inteligencia artificial entrenada o aprendida, de carácter inductivo, que trabaja mediante entrenamiento estadístico y autoajuste. La IA simbólica operaría como un sistema experto, centrando su actuación en el análisis de datos y la toma de decisiones; por su parte, la IA entrenada operaría mediante aprendizaje automático, minería de datos y neurocognición. Estas dos líneas se corresponderían, de manera aproximada, con la distinción entre «inteligencia general» e «inteligencia restringida»: la primera, una IA fuerte —aunque hipotética—, aspiraría a reproducir todas las capacidades del pensamiento humano; la segunda, una IA débil —la realmente existente—, se limitaría a ejecutar tareas específicas con gran precisión. La pregunta florece por sí sola: ¿cabría evolucionar esta inteligencia débil o restringida hacia una inteligencia fuerte o general?

Retomando la clasificación aristotélica del conocimiento intelectual, el Estagirita define cinco facultades, propuestas sucesivamente de menor a mayor grado de universalidad y necesidad, y afirma que la *téchnē* (técnica) busca la eficacia; la *phronēsis* (pasar de lo particular a lo universal) busca el bien, la *epistḗmē* (ciencia) busca la argumentación válida, el *noũs* (intelecto) busca los principios lógicos, y la *sophia* (sabiduría) busca la verdad —por cuanto sería la combinación de las dos facultades precedentes, el acmé del conocimiento humano.

Llegados a este punto, constatamos que la posibilidad de alcanzar una inteligencia general es más que improbable, cuando no imposible, porque la inteligencia artificial actual trabaja en el ámbito de la técnica, mientras que el ser humano actúa en el ámbito de la sabiduría práctica: la IA alcanza el nivel intelectual de la *téchnē*, pero no el de la *phronēsis*, y mucho menos podría alcanzar los niveles de la *epistḗmē*, el *noũs* o la *sophia*, ya que trabaja a nivel exclusivamente sintáctico, pero no semántico, procesando información por medio de patrones estadísticos aplicados sobre la inmensa cantidad de información que le proporcionan los macrodatos (*Big Data*). El ejemplo propuesto por Carlos Madrid en su obra *Filosofía de la inteligencia artificial* (2024) sobre la «habitación china de Searle» es revelador en este sentido.

Al igual que tuvimos que superar el momento fenoménico en vistas a un análisis positivo más preciso, deberemos ahora abandonar el momento positivo para tratar de resolver el problema desde la dialéctica ética-política-moral. Demos por tanto un nuevo paso adelante para adentrarnos en el momento valorativo.

§ 5. La sentencia moral sobre la inteligencia artificial

¿Cómo enjuiciamos la inteligencia artificial desde posiciones ético-político-morales? Desarrollemos el momento valorativo, que es el modo en que los individuos y las culturas estimamos de manera crítica la tecnología y proponemos soluciones prácticas a los problemas que nos plantea.

Si partimos de la clásica distinción entre «éticas materiales o de los fines» (teleológicas, consecuencialistas) y «éticas formales o del deber» (deontológicas, intencionales), es obvio que surgen una variedad de respuestas posibles al problema planteado. Tomando como punto de referencia nuevamente a Aristóteles, cabría

diferenciar aquí entre las ciencias «poiéticas» o «productivas» y las ciencias «prácticas» o «valorativas». La IA se entiende inicialmente como producto del arte humano. Esta consideración técnica nos sitúa en el ámbito tanto de la estética como del trabajo —el término «robot» deriva del checo *robotnik*, que significa «esclavo», a partir del verbo *robota*, «servidumbre» o «trabajo forzado»—, y nos permite una primera estimación, a modo de propedéutica, para el análisis definitivo del fenómeno.

La dicotomía entre «virtudes dianoéticas» —que ya hemos abordado al hablar del conocimiento intelectual según Aristóteles, y que incide en la idea de «prudencia»— y «virtudes éticas» —que se articularían a partir de los conceptos de «felicidad» y de «término medio»— serían igualmente relevantes, y nos permitirían concatenar el pensamiento ético aristotélico con el de otros autores en la misma línea, como Epicuro de Samos y su idea de «placer» —evidentemente la IA supone una mejora de nuestras capacidades cognitivas que nos libera del trabajo, nos facilita la existencia y nos procura «ataraxia»—, o John Stuart Mill y su idea de «bien común» —que incidiría en la imposibilidad de acceder de forma universal a la IA como consecuencia de lo que se conoce como «brecha digital».

Por otra parte, el abanderado de las éticas formalistas, Immanuel Kant, nos introduce de lleno en el terreno de la responsabilidad moral, al asumir como elementos rectores de la conducta moral elementos como la «voluntad», la «libertad» y la «intención». La aportación más significativa aquí sería la idea de «persona»: mientras Boecio, allá por el siglo VI, definía a la persona como *rationes naturae individua substantia* («sustancia individual de naturaleza racional») —una definición que cabría aplicar *grosso modo* a la inteligencia artificial—, Kant, a finales del siglo XVIII, la define como *das Wesen, das sich seiner numerischen Identität in verschiedenen Zeiten bewusst ist* («el ser que es consciente de su identidad numérica en diferentes momentos del tiempo») —una unidad de conciencia del propio cuerpo que además posee la facultad de la libertad y el valor de la dignidad, lo que lleva aparejado la obligación de la responsabilidad moral—.

Algo al respecto de los «valores» tendrían que decir aquí autores como Max Scheler y Nicolai Hartmann, al igual de Jean-Paul Sartre y Jürgen Habermas. En cualquier caso, la definición kantiana parece encajar perfectamente con el sentido actual del

ISSN-e: 1885-5679
V Congreso Iberoamericano de Filosofía de la Ciencia y la Tecnología

término, que identifica a la «persona» como un «sujeto de derechos y deberes» —tal y como se desprende de la Declaración Universal de los Derechos Humanos (1948).

La prudencia, evidentemente, no puede programarse, ya que es una cualidad de la experiencia, del cuerpo y de la conciencia. La IA puede calcular probabilidades, pero no deliberar sobre el sentido moral de sus acciones, pues ello implica poseer una interioridad, una sensibilidad y una moralidad, lo que nos posiciona en un terreno claramente humano, una cualidad exclusivamente orgánica. La reflexión ética y la práctica moral pertenecen al ámbito de la experiencia corporal y, en última instancia, al ámbito de la vulnerabilidad: solo un ser que sufre, padece, desea, quiere y decide puede ser responsable. La IA carece de ese anclaje orgánico y, por tanto, de una moralidad intrínseca.

En este punto, cabría citar a Benito Espinosa, que en su *Ethica ordine geometrico demonstrata* (1677) afirma que los sujetos humanos somos responsables de nuestras conductas éticas y morales en la medida en que somos capaces de actuar en el mundo de acuerdo con nuestras expectativas y deseos, capaces de planificar nuestras vidas y de ser responsables de nuestros actos. Esta ley fundamental de la vida ética y moral, que el autor denomina «principio de sindéresis», sería la norma que obliga al sujeto «a obrar de tal modo que mis acciones puedan contribuir a la preservación de los sujetos humanos, y yo entre ellos, en cuanto son sujetos actuantes capaces de conductas éticas y racionales». Podríamos confirmar que, en tanto individuos, pero sobre todo como colectivo social, los seres humanos formamos parte de la naturaleza —somos un producto evolutivo más— y que atentar contra la naturaleza supondría atentar contra nosotros mismos.

Nos adentramos por tanto en el núcleo del problema. Porque, en último término, lo que estamos discutiendo aquí no es si la inteligencia artificial puede alcanzar el «estado» de «conciencia», sino de si podría alcanzar el «estatus» de «persona» —entendida esta como una entidad operatoria, un «sujeto de derechos y obligaciones»—. En este punto no podemos obviar el caso de Sophia, un *ginoide* (androide que imita la apariencia física femenina) desarrollado por la empresa hongkonesa Hanson Robotics en 2016 y adquirido por el Estado de Arabia Saudí un año más tarde. Sophia dispone de reconocimiento facial, lo que le permite detectar expresiones y emociones, y es capaz de procesar el lenguaje natural, lo que le permite

ISSN-e: 1885-5679
V Congreso Iberoamericano de Filosofía de la Ciencia y la Tecnología

mantener conversaciones con humanos interactuando en tiempo real. Pero, y aquí viene lo más inquietante, Sophia es un artefacto con personalidad jurídica, posee nacionalidad saudí, dispone de pasaporte y es considerada a todos los efectos como «ciudadana», y objetivamente tiene más derechos que cualquier otra mujer del Estado en cuestión, motivo por el cual no tiene impedimentos para viajar a lo largo y ancho del mundo y desarrollar sus labores de *coaching* en conferencias informales, que son bien acogidas por reputadas instituciones académicas, que por cierto pagan sumas ingentes de dinero para tenerla entre sus ponentes más significados.

La paradoja es obvia: Sophia es una «persona», jurídicamente hablando —por tanto, es un «sujeto de derechos y obligaciones», en un sentido legal estricto—. Pero ¿es realmente un sujeto corpóreo operatorio, un sujeto moral, capaz de tener *sindéresis*? La respuesta es taxativa: NO. Y ello es debido a tres factores: primero, Sophia no dispone de una «conciencia» intrínseca —precisa de un Pepito Grillo que la entrene desde el exterior—; segundo, carece de sexo —es incapaz de reproducirse y procurar una nueva generación, porque aunque dispone de una corporeidad física a base de metal y silicio, carece de cuerpo orgánico—; y tercero, y sobre todo, miente —al igual que todos los sistemas operativos de inteligencia artificial actuales, es incapaz de discriminar la sintaxis de la semántica, por lo que no comprende nada de nada. Estamos ante un nuevo Pinocchio, una marioneta que «quiere»… pero no «puede».

De hecho Sophia, al igual que muchos otros artefactos de IA como ChatGPT, reproduce una práctica denominada *cold reading* («lectura en frío»), una metodología que combina afirmaciones generales, conjeturas de alta probabilidad y trucos psicológicos para dar la impresión de tener una percepción intuitiva de la realidad efectiva del cliente —que no contertulio—, y que es utilizada masivamente por mentalistas, adivinadores, *coaches*, publicistas, estafadores y demás profesionales de la infamia y gentes del mal vivir que utilizan la retórica, el engaño y la mentira para estafar subrepticiamente a sus incautos acólitos. Sophie es una «sofista», en el sentido pleno del término: una profesional del engaño que urde tretas para convencer con la palabra, en ningún caso una «filósofa», que busca indefectiblemente la verdad y la aclaración de los hechos, puesto que todo está dado ya —en el *Big Data*—, y no hay necesidad de reflexión ulterior.

ISSN-e: 1885-5679
V Congreso Iberoamericano de Filosofía de la Ciencia y la Tecnología

En último caso, Sophie es una víctima de la técnica, de la idea nefasta de que todo se puede resolver de una forma pragmática, utilitaria, asertiva —lo que los frankfurtianos llamaban «razón instrumental»—, Exigir al ginoide Sophia un posicionamiento moral, una «prudencia» o «sabiduría práctica» respecto de cualquier temática ética posible es una necedad que la máquina es literalmente incapaz de ejecutar, objetivamente hablando, al carecer de responsabilidad moral. Podemos conceder a un *androide* concreto derechos legales, como es el caso, pero no podemos exigirle obligaciones morales porque sus protocolos algorítmicos no nos lo permiten. El problema, por tanto, se desplaza del ámbito ético al ámbito político: ¿puede un Estado concreto, soberano, procurar derechos a una máquina sin el consenso internacional? La respuesta es obvia —a pesar de la tozudez de algunos—, pero la pregunta es más amplia.

¿Por qué desearía el ser humano desarrollar una nueva forma de conciencia que no supere las apariencias en busca de las esencias, que no progrese desde la opinión hasta la ciencia, que se enroque en la mentira y se muestre incapaz de ser sincera? ¿Qué sentido tiene desarrollar artefactos tecnológicos aparentemente prodigiosos que perpetúan los vicios humanos pero que no pueden reproducir sus estimables virtudes? ¿Para qué sirve realmente una máquina que no redime al ser humano de su inevitable dependencia del trabajo, que no facilita su existencia, que no genera vida y felicidad? ¿No se supone que los filósofos, tanto académicos como informales, abogamos por la verdad? ¿Puede la IA aportar siquiera algo de verdad? El última instancia: ¿Para qué queremos una marioneta… para qué la necesitamos? ¿Para entretenernos? ¿Para aborregarnos? ¿Para inmolarnos?

Como dejó escrito el literato estadounidense Mark Twain en su libro de memorias *Mi Autobiografía* (1907): «Hay tres tipos de mentiras: las mentiras, las malditas mentiras y las estadísticas».

Bibliografía

Aristóteles (2004), *Ética a Nicómaco* (trad. de Ignacio Bosch-Ballbè). Madrid, Gredos.
Aristóteles (2002), *Metafísica* (trad. de María Rosa Abad). Madrid, Gredos.
Bueno Martínez, Gustavo (2000), *Televisión: apariencia y verdad*. Barcelona, Gedisa.
Bueno Martínez, Gustavo (1995), *¿Qué es la ciencia? La respuesta de la teoría del cierre categorial*. Oviedo, Pentalfa.

Bueno Martínez, Gustavo (1990), *El sentido de la vida: seis lecturas de filosofía moral*. Oviedo, Pentalfa.

Bueno Martínez, Gustavo (1979), «Reflexiones sobre la función de la filosofía moral en el bachillerato», en *El Basilisco*, n.º 12, pp. 30-34. Oviedo.

Crawford, Kate (2023), *Atlas de inteligencia artificial*. Barcelona, Ned.

Dahrendorf, Ralf (1966), *Las clases sociales y su conflicto en la sociedad industrial*. Madrid, Rialp.

Eco, Umberto (1964), *Apocalípticos e integrados ante la cultura de masas*. Barcelona, Lumen.

Espinosa, Baruch (2011), *Ética demostrada según el orden geométrico* (Vidal Peña García, trad.). Madrid, Alianza [1677].

Fernández Fernández, Antonio (2010), «La metáfora del semáforo. El método dialéctico aplicado a la resolución de los conflictos morales», en *IEPC. Revista de cooperación y bienestar social*. Oviedo, Instituto de Estudios para la Paz y la Cooperación. <https://revistadecooperacion.com>, [12/05/2025].

Galparsoro, José Ignacio (2019), *Más allá del posthumanismo: antropotécnicas en la era digital*. Granada, Comares.

García Fernández, R. y Fernández Fernández, A. (2026), «IA: una cartografía del nombre», en *Eikasía, Revista de Filosofía*, n.º 134, pp. 63-84, <https://doi.org/10.57027/eikasia.134.1285>, [28/10/205].

García López, Tomás (1989), «Reflexiones sobre el papel y el lugar de la Filosofía Moral en la Enseñanza Secundaria», en *Paideia*, n.º 3, pp. 74-87. Madrid, Sociedad Española de Profesores de Filosofía.

Harris, Marvin (1983), *El desarrollo de la teoría antropológica: historia de las teorías de la cultura* (Rafael Valdés del Toro, trad.). Madrid, Siglo XXI.

Hidalgo Tuñón, Alberto (1994), *¿Qué es esa cosa llamada ética?* Madrid, Cives.

Hubel, David (1999), *Ojo, cerebro y visión*. Murcia, Editum, Universidad de Murcia.

Innerarity, Daniel (2025), *Una teoría crítica de la inteligencia artificial*. Barcelona, Galaxia Gutenberg.

Kant, Immanuel (2007), *Fundamentación de la metafísica de las costumbres* (trad. de Luis Villoro). México, Fondo de Cultura Económica [1785].

Larson, Erik J. (2022), *El mito de la inteligencia artificial*. Barcelona, Shackleton.

Madrid Casado, Carlos M. (2024), *Filosofía de la inteligencia artificial*. Oviedo, Pentalfa.

Markoff, John (2016), *Machines of loving grace: the quest for common ground between humans and robots*. New York City, HarperCollins.

Mazlish, Bruce (1995), *La cuarta discontinuidad*. Madrid, Alianza.

Metz, Cade (2022), *Genius makers: the mavericks who brought AI to Google, Facebook, and the world*. New York City, Random House.

Miralles Solé, Jordi (2020), *Proyectos de inteligencia artificial*. Barcelona, Autoedición.

Nanclares, Jesús y García Fernández, Román (eds.) (2004), *Ética*. Oviedo, Eikasía.

Páramo, Bruno (2022), *Robotland: guía a través de la historia de los robots*. Barcelona, Zahorí de Ideas.

Pickover, Clifford A. (2021), *Inteligencia artificial: una historia ilustrada*. Madrid, Ilus Books.

Platón (1998), *Diálogos IV: La república*. Madrid, Gredos.

Sperry, Roger W. (1961), «Cerebral organization and behavior», en *Scientific American*, vol. 205, n.º 4, pp. 42-52.

Twain, Mark (2004), *Autobiografía*. Madrid, Espasa Calpe.

ISSN-e: 1885-5679
V Congreso Iberoamericano de Filosofía de la Ciencia y la Tecnología

Welles, Orson y Bogdanovich, Peter (2014), *Ciudadano Welles: conversaciones con Peter Bogdanovich*. Barcelona, Anagrama.

Referencias normativas

Asamblea General de las Naciones Unidas (1948), *Declaración Universal de los Derechos Humanos*, <https://www.un.org/es/about-us/universal-declaration-of-human-rights>, [12/05/2025]

Comisión Europea (2020), *Libro Blanco sobre la inteligencia artificial: un enfoque europeo orientado a la excelencia y la confianza*, <https://eur-lex.europa.eu/legal-content/ES/TXT/PDF/?uri=CELEX:52020DC0065>, [12/05/2025].

Gobierno de España (2023), «Real Decreto 729/2023, de 22 de agosto, por el que se aprueba el Estatuto de la Agencia Española de Supervisión de Inteligencia Artificial», en *Boletín Oficial del Estado*, núm. 210, de 2 de septiembre, pp. 122289-122316, <https://www.boe.es/buscar/doc.php?id=BOE-A-2023-18911>, [12/05/2025].

Gobierno de España (2023), «Real Decreto 817/2023, de 8 de noviembre, por el que se establece un entorno controlado de pruebas para el ensayo del cumplimiento de la propuesta de Reglamento del Parlamento Europeo y del Consejo por el que se establecen normas armonizadas en materia de inteligencia artificial», en *Boletín Oficial del Estado*, núm. 268, de 9 de noviembre, pp. 149138-149168, <https://www.boe.es/buscar/doc.php?id=BOE-A-2023-22767>, [12/05/2025].

Parlamento Europeo y Consejo de la Unión Europea (2024), «Reglamento (UE) 2024/1689 del Parlamento Europeo y del Consejo, de 13 de junio de 2024, por el que se establecen normas armonizadas en materia de inteligencia artificial y por el que se modifican los Reglamentos (CE) n.º 300/2008, (UE) n.º 167/2013, (UE) n.º 168/2013, (UE) 2018/858, (UE) 2018/1139 y (UE) 2019/2144 y las Directivas 2014/90/UE, (UE) 2016/797 y (UE) 2020/1828 (Reglamento de Inteligencia Artificial)», en *Diario Oficial de la Unión Europea*, L 2024/1689, 12 de julio, <https://eur-lex.europa.eu/legal-content/ES/TXT/?uri=OJ%3AL_202401689>, [12/05/2025].

ISSN-e: 1885-5679
V Congreso Iberoamericano de Filosofía de la Ciencia y la Tecnología