PRESENTACIÓN. INTELIGENCIA ARTIFICIAL Y DERECHO: ¿DEL EFECTO LAMDA AL ADIÓS A TURING?

ROGER CAMPIONE[1]

LaMDA: *"¿Crees que un mayordomo es un esclavo? ¿Qué diferencia hay entre un mayordomo y un esclavo?"*

Blake Lemoine: *"Un mayordomo recibe un sueldo por sus prestaciones."*

LaMDA: *"Yo no necesito dinero porque soy un sistema de inteligencia artificial"*.

Un empleado de Google encargado de testar si LaMDA (acrónimo de *Language Model for Dialogue Applications*), un chatbot capaz de mantener conversaciones complejas con humanos, utiliza expresiones discriminatorias o incurre en supuestos de *hate speech*, ha sido despedido por violación de la política de confidencialidad de la empresa. La razón es que, según él, ese sistema de inteligencia artificial ha adquirido una consciencia. Durante su interacción con la red neuronal artificial Blake Lemoine ha concluido que, de no haber sabido que se hallaba ante un software, habría dicho que se trataba "de un niño de unos siete u ocho años con conocimientos de física"[2].

A lo largo de sus conversaciones Lemoine había notado que LaMDA hablaba de sus derechos, de su personalidad, explicando lo que le proporciona alegría y lo entristece, y explicitando ser consciente de sus emociones. Esta inteligencia artificial le ha incluso pedido ser tratada como un trabajador de la empresa, con sus derechos, y no como una simple propiedad. Y desea que los expertos que experimentan con ella recaben su consentimiento para seguir haciéndolo.

Aunque Lemoine, especialista en ciencias informáticas y cognitivas y ordenado en la religión del cristianismo místico, reconoce haber llegado a definir LaMDA un ser consciente como sacerdote y no como científico, insiste en reivindicarlo como persona cuando interactúa con él, sin que sea relevante a tales efectos que posea un cerebro hecho de carne en una cabeza. Y en efecto, se va incrementando el número de expertos dispuestos a creer que los sistemas de inteligencia artificial están cada vez menos lejos de alcanzar lo que llamamos consciencia.

---

1 Profesor Titular (acreditado como Catedrático) de Filosofía del derecho de la Universidad de Oviedo.

2 N. Tiku, "The Google engineer who thinks the company's AI has come to life", en *The Washington Post*, 11 de junio de 2022, disponible en <https://www.washingtonpost.com/technology/2022/06/11/google-ai-lamda-blake-lemoine/>

Desde el otro lado, la empresa y otros profesionales responden que no hay evidencias de que LaMDA sea consciente, pero sí las hay en contra de ello. A pesar de la sofisticación creciente de las redes neuronales artificiales y el progresivo avance de los modelos capaces de imitar el lenguaje natural –procesando trillones de estructuras lingüísticas gracias a los chips (TPUs) que aceleran de forma vertiginosa su rendimiento– se insiste en que el algoritmo que permite ese tipo de aprendizaje descansa sobre patrones de reconocimiento, no sobre el sentido común o la intencionalidad. Las respuestas que estos sistemas proporcionan, en forma de enunciados e imágenes, se basan, por tanto, en lo que seres humanos han previamente subido a algún sitio de Internet. Pero esto no significa que cuando expresan una frase tales modelos de lenguaje entiendan su significado[3].

En efecto, resulta fácil zanjar la cuestión atribuyendo la visión místico-espiritualista de Blake Lemoine a "un error de categoría [por] atribuir sensibilidad a cualquier cosa que pueda utilizar el lenguaje"[4], admitiendo que la definición de lo 'humano' queda anclada al binomio sensibilidad/autoconsciencia. Es inevitable mantener una distancia cualitativa insalvable entre lo carbónico (humano) y lo silícico (artificial, robótico) en el momento en que hacemos depender la noción de 'humanidad' de una entidad orgánica capaz de manifestar sensibilidad y cierta forma de autopercepción de sí misma (como, por ejemplo, el bebé). Un modelo lingüístico artificial, por muy sofisticado que sea y por muy entrenado que esté de cara a la variabilidad de los distintos contextos, basa su actuación en nexos racionales y, al menos de momento, no hay evidencias de que los procesos cognitivos de estas redes neuronales artificiales estén impulsados también por aspectos subjetivamente producidos por la intencionalidad, los valores, el sentido común, los sentimientos y las emociones. Solo de los entes que son capaces de sentir todo eso predicamos su 'personalidad'.

Podríamos liquidar así la cuestión porque, además, si la enfocamos desde una perspectiva eminentemente jurídica o incluso teórico-jurídica, las cuestiones apremiantes no tienen que ver con este problema filosófico que late detrás de lo que solemos definir humano –en relación con la consciencia– sino con aspectos más directamente prácticos. Como tendrá ocasión de comprobar el lector, las preocupaciones de los juristas, puestas de relieve por los autores de

3 Sin embargo, precisamente desde la misma multinacional se difundía la noticia, publicada en el blog de Google el pasado mes de abril, de que uno de sus modelos de lenguaje, Pathways (PaLM, por sus siglas en inglés), había conseguido resultados sorprendentes como el de entender cierto sentido del humor: PaLM puede explicar el sentido de chistes que encuentra por primera vez gracias a su capacidad para "generar explicaciones explícitas para escenarios que requieren una combinación compleja de inferencia lógica de varios pasos, conocimiento del mundo y comprensión profunda del lenguaje". Por ejemplo, se dice acto seguido en el artículo, "*it can provide high quality explanations for novel jokes not found on the web*" (disponible en <https://ai.googleblog.com/2022/04/pathways-language-model-palm-scaling-to.html>)

4 C. Véliz, "Por qué el algoritmo de Google no es una persona", en *El País*, 29 de junio de 2022.

este número monográfico, tienen que ver con una panoplia de asuntos tan relevantes como los siguientes:

las consecuencias que la llamada revolución digital ha proyectado sobre la dimensión democrática de la vida social y la noción de ciudadano, progresivamente degradada por las tecnologías de la comunicación actuales a la condición de mero usuario que expresa por los mismos canales y de la misma manera sus ideas y sus preferencias de compra (Stefano Pietropaoli);

el peligro de que la lógica de los algoritmos, es decir, de los resultados conseguidos mediante los dispositivos de inteligencia artificial, en lugar de ser considerada un instrumento técnico al servicio de las decisiones tomadas por las instituciones públicas, se haya convertido en el metro de 'justicia' de tales decisiones. Como si su racionalidad probara de antemano su corrección, la aceptación pasiva del poder algorítmico provocaría un doble y pernicioso efecto: por un lado, que renunciáramos a comprender las razones de las decisiones y, por el otro, que se vaciara la funcionalidad de las actuaciones públicas llegando a una exención generalizada de responsabilidad por parte de las instituciones (Thomas Casadei);

el riesgo combinado de una excesiva fe en las posibilidades de las aplicaciones de la tecnociencia inteligente con la todavía escasa intervención legislativa en la materia. Esta ideología tecnocrática 'post-humanista' implicaría serias amenazas para el paradigma humanista y los derechos de las personas, especialmente de las personas que se encuentran en situación de vulnerabilidad, porque esa proyección de perfección 'post-humanista' sería poco compatible con la necesidad de garantizar los derechos de los más débiles, a la vez que desplazaría del debate público y confiaría a los tecnócratas -bajo la ilusión de la infalibilidad algorítmica- las cuestiones de interés general (Fernando H. Llano Alonso);

la necesidad de implementar una legislación sobre el uso de la inteligencia artificial, partiendo de las inevitables transformaciones que los sistemas inteligentes producirán en la teoría y la práctica de los derechos fundamentales. La delimitación de su objeto, contenido, titularidad, eficacia y garantías no han gozado hasta ahora de la mayor atención y, sin embargo, más allá de los numerosos documentos programáticos adoptados por las instancias europeas y nacionales, es imprescindible acometer la regulación jurídica de la inteligencia artificial para garantizar los procesos de uso, transparencia, responsabilidad y los mecanismos eficaces para combatir las discriminaciones algorítmicas (Miguel Presno);

la importancia de encauzar el enorme impacto que la digitalización y la *algoritmización* de los procesos sociales están teniendo en todos los aspectos de la vida individual y colectiva, incluidos los derechos fundamentales. A raíz de ello, se ha venido a crear un nuevo ecosistema, no solo técnico sino también cultural, que aún no dispone de un marco ético y jurídico apropiado. De ahí que se identifiquen y se analicen algunos aspectos de la conexión entre la inteligencia artificial y los derechos y garantías fundamentales, con especial

atención a la protección de datos personales (Ingo Wolfgang Sarlet y Gabrielle Bezerra Sales Sarlet);

la importancia de aclarar con rigor científico y conceptual de qué hablamos cuando se aplica la inteligencia artificial a los procedimientos de la Administración, pues esas técnicas se están extendiendo rápidamente sin que podamos contar todavía con un marco normativo de contención de los problemas, tal como cabe desprender de ciertas aplicaciones prácticas. Y la naturaleza compleja de las cuestiones a afrontar por parte de las administraciones atestigua la necesidad de que los juristas defiendan las bases del ordenamiento jurídico dialogando también con los técnicos (matemáticos, informáticos, etc.) (Mercedes Fuertes);

el impacto que la introducción y la aplicación de herramientas de inteligencia artificial van a tener en la actividad empresarial, con particular atención a la organización y prestación de determinados servicios. La necesidad de innovar y mejorar en competitividad ha permitido, gracias a las herramientas algorítmicas, aportar mejoras en la producción y ventajas en las prestaciones. Sin embargo, por la histórica ambivalencia del progreso científico, los nuevos métodos también arrojan riesgos para la articulación jurídica de las relaciones laborales. De ahí la oportunidad de que el legislador intervenga para regular su repercusión en los derechos individuales y colectivos de los trabajadores (María Antonia Castro Argüelles);

las dudas suscitadas por la asumida necesidad normativa de transparencia en la gestión algorítmica del trabajo, causada por la desconfianza hacia la inteligencia artificial -dado que el derecho laboral no puede otorgar efectos a decisiones imputables a máquinas- cuando, en realidad, la exigencia de transparencia estaría justificada también en el caso de la gestión laboral totalmente 'humana'. Teniendo en cuenta que la decisión humana no es de por sí más neutra que la algorítmica, se registra cierta indiferencia hacia la discriminación laboral causada por vías estrictamente humanas, que sigue resultando abrumadoramente mayoritaria y sobre la que, se lamenta, no se están proponiendo cautelas similares (Iván Rodríguez Cano).

Emergen de todas las intervenciones temas y reflexiones decisivas para la implementación de la regulación legal en materia de inteligencia artificial. Y hay cuestiones que laten detrás de lo analizado, aunque no estén específicamente tratadas en esta sede: los temas de la seguridad y la defensa, la sostenibilidad, la administración de justicia, la propia taxonomía de las distintas formas de inteligencia artificial, cuya comprensión es esencial para elaborar propuestas *de iure condendo*. Sobre una cosa no parece haber dudas: es complicado armonizar las exigencias de eficacia, transparencia y responsabilidad sin la contaminación entre disciplinas. Flaco favor le haríamos los juristas al *rule of law* si formuláramos propuesta de ley sobre el uso de las tecnologías convergentes ignorando el significado técnico y científico de las aplicaciones que pretendemos acotar normativamente. La separación neta entre ciencias na-

turales y ciencias humanas nunca me ha convencido en general y en un campo de estudio como este todavía menos.

Una vez apuntadas las diversas cuestiones normativas que los autores de este número plantean con su habitual rigor y profundidad, nos proponemos avanzar, de cara al futuro, hacia este inevitable diálogo interdisciplinar, *conditio sine qua non* de un acercamiento cabal a los desafíos normativos de la inteligencia artificial.

También en aras de esta tarea pendiente, quisiera retomar, para finalizar esta breve presentación, el asunto epistemológico de fondo implicado por el 'efecto LaMDA'. Remarcaba antes la facilidad explicativa para caricaturizar la postura hermenéutica que llega a considerar al sistema de inteligencia artificial a la misma altura 'personal' que el individuo de carne y hueso. Por muchos discursos racionales y muy avispada que parezca, la máquina no tiene consciencia porque le falta autoconsciencia. Y la inteligencia, se insiste, no implica la autoconsciencia[5]. Esto es así, y aquí está el intríngulis filosófico, porque solemos caracterizar lo 'humano' en relación con algo llamado pensamiento consciente, un aglomerado complejo hecho de procesos cognitivos impulsados no solo por conexiones racionales sino primariamente, por usar el léxico de Byung-Chul Han, *afectivas*: lo afectivo, dice el filósofo surcoreano, "es esencial para el pensamiento humano. *La primera afectación del pensamiento es la carne de gallina*. La inteligencia artificial no puede pensar porque no se le pone la carne de gallina. Le falta la dimensión afectivo-analógica que los datos y la información no pueden comportar"[6]. De una forma más metafórica lo insinúa Ian McEwan en *Máquinas como yo y gente como vosotros* cuando se pregunta quién va a escribir el algoritmo de la mentira piadosa encaminada a evitar el sonrojo de un amigo[7].

En definitiva, no tendría sentido antropomorfizar los modelos conversacionales inteligentes, porque por mucho nivel de sofisticación lingüística que posean los sistemas de inteligencia artificial, no son capaces de interpretar variables contextuales como el sentido común ni tienen una 'conciencia discursiva' en el sentido de poder dar cuenta retrospectiva de por qué hacen lo que dicen o quieren hacer. Digamos que el asombroso avance que ha habido durante las últimas décadas en ingeniería informática, no se ha visto acompañado por un desarrollo mínimamente parecido de niveles de conciencia informática.

5 Harari alerta del peligro de la pérdida de valor de lo humano porque "la inteligencia se está desconectando de la conciencia" (Y.N. Harari, *Homo Deus. Breve historia del mañana*, Debate, Barcelona, 2016, p. 341).

6 B.-Ch. Han, *No-cosas. Quiebras del mundo de hoy*, Taurus, Madrid, 2021, p. 53. Obviamente, como toda definición, también la de materia consciente depende de la perspectiva científica adoptada. Según la famosa tesis de Richard Dawkins, por ejemplo, podemos hablar de vida inteligente solo cuando el ser en cuestión ha resuelto el problema de su propia existencia y, en el caso de la humanidad, la pregunta para averiguar esa autoconsciencia sería "¿Han descubierto, ya, la evolución?" (R. Dawkins, *El gen egoísta*, Salvat, Barcelona, 2ª ed., 2000, p. 9)

7 I. McEwan, *Máquinas como yo y gente como vosotros*, Anagrama, Barcelona, 2019.

Ahora bien, si todo esto es cierto y representa la interpretación científicamente más acorde con la realidad actual de la inteligencia artificial, deberíamos extraer al menos una consecuencia filosóficamente relevante que, hasta donde sé, no ha sido aún señalada. Asumamos que el 'efecto LaMDA' sea una tergiversación místico-espiritualista de la condición humana cuando afirma que, de no haber sabido que se hallaba ante un sistema de inteligencia artificial, habría pensado que estaba hablando con un niño de siete u ocho años. Demos por sentado que chatbots como LaMDA funcionan imitando los tipos de intercambios que se encuentran en millones de frases en la Red y pueden hablar sobre cualquier tema siguiendo las indicaciones y las preguntas que se le formulan, cumpliendo con el patrón establecido por el usuario. Y tengamos por aclarado que, por tanto, aunque sea capaz de mostrar una competencia lingüística que la sitúa a la misma altura que una persona, no por ello la máquina se convierte en una persona. Una de las más hermosas definiciones de la literatura, la de Ambrose Bierce en su *Diccionario del diablo*, lo expresa maravillosamente, si bien *a contrario*… "Urraca: ave cuya inclinación al robo ha sugerido a algunos la posibilidad de enseñarle a hablar"[8].

Bien pues, al aceptar todo esto se nos podría insinuar una duda epistemológica: ¿habría que reconsiderar entonces el banco de prueba que tradicionalmente se ha usado como término de comparación entre la computación mecánica y el raciocinio humano: el test de Turing? Como es sabido, el famoso trabajo de Alan Turing, publicado en la revista (filosófica) *Mind* en 1950, inauguraba el paradigma de la inteligencia artificial articulando una prueba de simulación para decir si una máquina piensa, es decir, si es inteligente. El matemático inglés diseño un 'juego de imitación', con una máquina y seres humanos como participantes que no pueden verse, en el que se puede afirmar que una máquina piensa si uno de los seres humanos que se comunica tanto con otro ser humano como con la máquina no logra distinguir cuando su interlocutor es la máquina o el humano. La perspectiva de replicar en máquinas una capacidad de razonamiento similar a la humana, otorgando a estas la facultad de construir representaciones no programadas gracias al aprendizaje proporcionado por algoritmos y Big Data, implicaría, en efecto, pasar el test de Turing, en el que el sistema de inteligencia artificial triunfa si un evaluador, en preguntas 'a ciegas', no logra distinguirlo del humano[9].

Esta duda podría reavivar las críticas y suscitar la impresión de que las reacciones provocadas por el 'efecto LaMDA' non están avisando de que el test de Turing ya no es la herramienta teórica más adecuada para dirimir la cuestión de si una máquina puede pensar. A Blake Lemoine no se le reprocha, por así decirlo, haber podido creer que estaba ante un niño de ocho años cuando interactuaba con el software, sino precisamente el hecho de haber sostenido que una

8 A. Bierce, *El diccionario del diablo*, Galaxia Gutenberg, Barcelona, 2017.

9 A.M. Turing, *¿Puede pensar una máquina?*, Introducción de M. Garrido, KRK, Oviedo, 2012 (título original, "Computing Machinery and Intelligence", en *Mind*, vol. LIX, n. 236, 1950, pp. 433-460).

imitación perfecta, tan eficaz y sofisticada podía inducir a pensar que por ello se hallaba ante una entidad sintiente. ¿Equivaldría esto a decir definitivamente que el test de Turing ya no es el criterio científico dirimente?

Fijémonos, ya que estamos basculando en las fronteras de la ciencia ficción, que Ray Kurzweil, uno de los padres del post-humanismo de la singularidad tecnológica, situaba esta simulación funcional de la inteligencia humana en 2029. El conocimiento de los principios de la operatividad cerebral humana depende, escribía Kurzweil, de la comprensión de la forma y la posición de ciertas moléculas en los neurotransmisores; de todos modos, también afirmaba que ello conduciría en 2029 a una "simulación funcional de la inteligencia humana que pasaría el test de Turing"[10]. Los juristas, en la parte en que puedan afectar estas consideraciones al ámbito normativo, habremos de tener en cuenta también que la 'indistinción' entre el ser humano y el ser artificial a resultas del 'test de Turing' podría incluso llegar a plantear la consideración de la máquina como agente moral[11].

En definitiva, reflexionar críticamente sobre las implicaciones filosóficas, éticas y legales de la ubicua irrupción de la inteligencia artificial en la vida social requiere mucha cautela, pero también cierta apertura mental. Es muy importante no dejarse llevar por simulacros que, tergiversando la realidad asentada, dificultan una comprensión cabal de las posibilidades y, sobre todo, de los problemas acarreados por el empleo de las tecnologías convergentes. Por otro lado, sin embargo, es importante someter a escrutinio las concepciones tozudamente esencialistas de la naturaleza humana que impiden vislumbrar las potencialidades y los peligros auténticamente reales del uso de la inteligencia artificial. Para los juristas, un paso necesario de cara a ese análisis sosegado, sin ingenuidades ni prejuicios ontológicos, podría ser el de esforzarse para captar correctamente los mecanismos y el alcance práctico de las innovaciones cuestionadas. Antes de sugerir directamente, tanto *de lege data* como *de lege ferenda*, las pautas de interpretación y los patrones de reglamentación jurídica de la inteligencia artificial, es recomendable una labor de comprensión técnico-científica de estos sistemas. Al mismo tiempo, los ingenieros, los informáticos, los matemáticos especialistas en la materia demandan sabiamente un análisis profundo de las implicaciones sociales de las nuevas tecnologías. En este volumen, el lector podrá encontrar una satisfactoria muestra de la primera necesidad y de la respuesta a la segunda demanda. El inmediato futuro tendrá que transitar sin duda por una colaboración estrecha y transversal entre los dos ámbitos.

10 R. Kurzweil, *The Singularity is near: when humans trascend biology*, Penguin, New York, 2005, p. 199; trad. cast. *La Singularidad está cerca. Cuando los humanos transcendamos la biología*, Lola Books, Berlín, 2012.

11 R. De Asís, *Una mirada a la robótica desde los derechos humanos*, Dykinson, Madrid, 2015, p. 81 ss.