De la herramienta a la agencia

Algunas alertas ante la irrupción de la IA

Luis Xavier López-Farjeat

La emergencia de la IA ha empezado a transformar la manera en que nos relacionamos con nosotros mismos, con el mundo y con nuestro quehacer en él. Esto ha generado una suerte de polarización entre los que López-Farjeat llama "tecnofetichistas" y "tecnopesimistas". En medio de esa polémica, López-Farejeat hace un llamado a la mesura y, mediante una revisión exhaustiva de autores que han pensado el tema de manera ponderada, enciende algunas alertas acerca de lo que, como humanidad, habremos de enfrentar ante esa extraña cosa —mitad herramienta, mitad tecno-cerbero— que, creada por el propio ser humano, llegó para quedarse con nosotros.

Un fenómeno tan disruptivo como la inteligencia artificial ha polarizado los puntos de vista: por una parte, los tecnoutopistas, cuya versión más extrema ha derivado en un tecnofetichismo; por otra, los tecnopesimistas y su visión apocalíptica. Un llamado a la mesura no estaría de más. La inteligencia artificial es como un animal: o lo domesticas o te come. Hay quienes piensan, como Javier Sicilia, que no habrá forma de domesticarla, que terminará por aniquilarnos. Otros, menos pesimistas, como León XIV, creen que, al igual que toda herramienta, habrá que aprender a usarla de la mejor manera en beneficio de la humanidad. En su encíclica *Magnifica Humanitas*, rechaza todo uso de la inteligencia artificial que reduzca a la persona a un dato, un perfil, un consumidor o un engranaje productivo. Añade que, cuando se habla de la ética de la inteligencia artificial, el asunto no puede centrarse en la exigencia de darle un "buen uso". La preocupación ha de estar presente desde el diseño de la herramienta. El problema, a mi juicio, es que la inteligencia artificial no es una herramienta cualquiera.

Heidegger entendió la herramienta como un elemento fundamental mediante el cual podemos relacionarnos con el mundo. Las herramientas, en ese sentido, dejan de resultarnos ajenas; adquieren un uso cotidiano. No nos resulta extraño encender cada mañana la cafetera, tomar un cuchillo para untar mermelada en una rebanada de pan; nos resulta familiar subir al auto, tomar las llaves, abrir la cerradura de la oficina y encender la computadora. Podríamos hacer un recuento de los artefactos que tenemos a la mano para su uso. Al empleo cotidiano de las herramientas, Heidegger le llamó *Zuhandenheit* ("ser a la

mano"). A su vez, las herramientas están inmersas en una red de referencias (*Verweisung*): cada mañana tomo el bote de café y, con la cucharita medidora, pongo la cantidad necesaria para el número de tazas que pienso tomar; vierto el agua en la cafetera y, cuando todo está listo, sirvo mi café. Todos estos artefactos forman parte de una red de propósitos y encuentran su propio sentido en su función. Si el cuchillo pierde su filo, la cafetera se descompone y el auto ya no enciende, pierden su función y se convierten en artefactos inservibles. Pasan a ser lo que Heidegger llama *Vorhandenheit* ("disponibilidad").

La filosofía analítica abomina la verbosidad heideggeriana. Hay que admitir, sin embargo, que en varios casos arroja significados útiles, aunque a menudo complejos. La diferencia entre *Zuhandenheit* y *Vorhandenheit* es un buen ejemplo: "estar a la mano" y "estar ahí adelante" o "estar a disposición". Los objetos que están a la mano son aquellos que usamos incluso de manera inconsciente. Se han vuelto una extensión de nosotros: usar un tenedor, escribir con un bolígrafo o con un teclado, usar el teléfono celular, etc. Cuando por algún motivo —una avería o la misma obsolescencia— los objetos dejan de ser útiles, dejan de estar "a la mano" y se vuelven simplemente cosas que tenemos "ahí delante", entonces los vemos como algo ajeno. Si estas categorías heideggerianas fuesen aplicables a nuestro entorno actual, ¿qué clase de herramienta sería la inteligencia artificial? ¿Habríamos de preocuparnos si admitiéramos que, al menos entre sus usuarios, ha empezado a convertirse en una "herramienta a la mano"?

Heidegger tenía en mente herramientas destinadas a facilitar el trabajo físico. Hannah Arendt distingue en *La condición humana* entre las actividades ligadas a la subsistencia (aquello que denomina "labor"), las actividades destinadas a construir el mundo artificial humano (aquello que llama "trabajo" y es precisamente la construcción de herramientas y artefactos), y aquellas actividades que se dan a través de la palabra y los hechos, sin la mediación de las herramientas (aquello que llama "acción" para referirse a las personas que participan en la creación del espacio político). La acción involucra aquello que nos caracteriza como seres humanos: el pensamiento y la palabra. Enciendo una alerta: la inteligencia artificial no es una herramienta que facilite el trabajo físico; más bien intenta emular nuestro lenguaje, nuestras lenguas y, por lo tanto, nuestro pensamiento. Diego Rosales lo ha visto con acierto en este mismo número de *Conspiratio*: la apropiación del pensamiento y el lenguaje por parte de la inteligencia artificial es fantasmagórica. Por decirlo de manera un tanto brutal, la inteligencia artificial no se limita a facilitar actividades manuales o en las que 'metemos el cuerpo'; pretende, más bien, simplificar nuestros procedimientos intelectuales.

La irrupción de la inteligencia artificial podría analogarse a la de la escritura tal como se relata en el *Fedro* de Platón. En *Fedro o la Escuela al Museo*, Iván Illich recuerda cómo, en el conocido mito sobre el origen de la escritura, Platón hace que Sócrates lo describa como un *pharmakon* para la humanidad, capaz de actuar según su uso, ya sea como veneno o como remedio. Así como la escritura puede ser un instrumento extraordinariamente útil, Sócrates la considera una prótesis del pensamiento y, en consecuencia, un riesgo para la memoria. El efecto de la inteligencia artificial es similar y más radical.

Escuché recientemente una conferencia de André Laks titulada "Plato on IA". Sin conocer el texto de Illich, Laks contó que, a través de un amigo, utilizó por primera vez la IA llamada *Claude*. Al preguntarle por su opinión sobre aquel pasaje en el que Thot presenta la escritura como un remedio para la memoria y la sabiduría, Claude le advirtió amablemente que no tomara este pasaje como una profecía, señalando que la similitud entre la reflexión de Platón sobre la escritura y nuestras preocupaciones actuales sobre la inteligencia artificial era solamente "estructural" ("analógica"), no material. No estoy tan seguro. Platón, bien lo recordaba Laks, admite un uso legítimo de la escritura: la filosofía puede utilizarla para recordar el contenido de su conocimiento. Si Platón no hubiese escrito sus *Diálogos*, habríamos perdido un legado intelectual invaluable. No parece del todo claro que la función de la inteligencia artificial se limite a preservar nuestras ideas. Enciendo otra alerta: en su caso hay un riesgo mayor, a saber, la sustitución de nuestra capacidad para pensar por nosotros mismos. Pero, ¿puede realmente la inteligencia artificial sustituir a la inteligencia humana?

Uno de los mejores libros que abordan esta cuestión es el de Erick J. Larson, *The Myth of Artificial Intelligence: Why Computers Can't Think the Way We Do* (2021). Larson sostiene que carecen de fundamento científico quienes creen que estamos en el camino correcto hacia una inteligencia artificial de nivel humano y que sólo hace falta más cómputo, más datos y más tiempo para alcanzarla. Según Larson, una inteligencia artificial equiparable a la humana no existe todavía, ni siquiera como programa técnico definido, sino más bien sólo como imaginación cultural, promesa empresarial y relato futurista. Los sistemas contemporáneos pueden ser extraordinariamente eficaces en tareas delimitadas: jugar al ajedrez o clasificar imágenes, detectar patrones, traducir textos, recomendar productos o responder ciertas preguntas. Pero para Larson esos logros no constituyen pasos acumulativos hacia la inteligencia humana. Son tan sólo éxitos en dominios acotados, formalizables y dependientes de grandes volúmenes de datos. El

error consiste en extrapolar esos triunfos locales a una inteligencia general capaz de comprender contextos abiertos, formular hipótesis, interpretar situaciones ambiguas y actuar con sentido común. La distancia entre reconocer patrones y comprender el mundo no sería gradual, sino cualitativa.

Lo que me resulta más convincente de la aproximación de Larson es su carácter epistemológico. La historia de la inteligencia artificial ha explorado, según explica, sobre todo dos formas de razonamiento: la deducción, propia de la inteligencia artificial clásica y simbólica, y la inducción, propia del aprendizaje automático y de la inteligencia artificial basada en datos. La deducción opera a partir de reglas y preserva la verdad de las premisas, mientras que la inducción generaliza a partir de regularidades observadas. Ambas son eficientes, pero insuficientes. La inteligencia humana, en contraste, depende de manera categórica de la abducción, es decir, de la capacidad de formular conjeturas plausibles para explicar hechos sorprendentes o ambiguos. Inspirado en Charles S. Peirce, Larson insiste en que pensar no se reduce a calcular o generalizar, sino que también implica adivinar bien, proponer hipótesis, leer indicios, revisar interpretaciones y moverse en un mundo en el que la información siempre es incompleta. Las máquinas no operan en el vacío, sino que necesitan guiarse por conjeturas humanas, por hipótesis inteligentes sobre mensajes, claves y patrones posibles. Dicho de otro modo, el éxito técnico de la inteligencia artificial descansa sobre una capa previa de interpretación abductiva: la inteligencia real no consiste simplemente en ejecutar procedimientos mecánicos, sino en establecer las condiciones iniciales pertinentes para que esos procedimientos tengan sentido.

¿Qué hace Claude, Grok, ChatGPT o cualquier otra inteligencia artificial cuando se le pide interpretar, como hizo Laks, un pasaje de los *Diálogos* de Platón? ¿Qué haría, por ejemplo, con *Tirano Banderas*, de Ramón del Valle-Inclán, una obra compuesta con un léxico complejísimo y un uso variado de americanismos? ¿Cómo interpretaría un pasaje bíblico o un pasaje coránico? Estos ejemplos requieren una cantidad considerable de recursos lingüísticos, gramaticales, filosóficos (teológicos, en algunos casos), históricos, contextuales, etc. Claude arrojará, sin duda, una explicación coherente —interesante y sugerente, incluso— construida a partir de las más variadas interpretaciones que los seres humanos hemos hecho. La inteligencia artificial parece comprender el lenguaje. Sin embargo, vuelvo a Larson cuando sostiene que comprender una lengua no equivale a manipular cadenas de símbolos ni a extraer correlaciones estadísticas. El lenguaje cotidiano —al igual que el filosófico y el literario, añadiría por mi cuenta— está lleno de ambigüedad, presupuestos implícitos, referencias contextuales, humor e ironía,

conocimiento del mundo y expectativas pragmáticas. El problema al interpretar un texto no es simplemente gramatical, sino que exige comprender qué es plausible en el mundo. Por eso, para Larson, el lenguaje natural revela con particular claridad los límites de una inteligencia artificial basada en la frecuencia, la correlación y los patrones superficiales.

Larson llama "kitsch tecnológico" a la simplificación sentimental y empobrecida con la que la cultura contemporánea transforma problemas filosóficos profundos en relatos fáciles sobre máquinas cada vez más inteligentes. El concepto es provocador: así como el kitsch reemplaza la complejidad estética o moral por una emoción prefabricada, la mitología de la inteligencia artificial sustituye preguntas difíciles sobre inteligencia, conciencia, creatividad, ciencia y sentido humano por una narrativa lineal de progreso inevitable. En esa narrativa, la técnica aparece como destino, y la imaginación humana queda subordinada a un futuro supuestamente ya escrito.

A pesar de su posicionamiento crítico, Larson no es antitecnología. Reconoce que la inteligencia artificial tiene usos valiosos y puede funcionar como una prótesis de la inteligencia humana, del mismo modo que el telescopio o el microscopio ampliaron nuestras capacidades perceptivas. Su objeción no se dirige contra las herramientas computacionales, sino contra su conversión en ideología. Los sistemas de inteligencia artificial pueden ayudar a clasificar, buscar, procesar, sugerir y detectar patrones. Una alerta más: el peligro surge cuando confundimos esas capacidades con la comprensión, la agencia o la inteligencia general. En ese momento, cedemos autoridad a máquinas que no entienden lo que hacen.

En *The Ethics of Artificial Intelligence* (2023), Luciano Floridi sostiene precisamente que la inteligencia artificial debe entenderse como una nueva forma de agencia, no como una herramienta ni como un sustituto de la inteligencia humana. Esto significa que la verdadera novedad de esa supuesta inteligencia no radica en hacer que las máquinas "piensen" como los seres humanos, sino en que pueden "actuar" eficazmente en determinados entornos sin requerir inteligencia en sentido humano. En vez de preguntar si la inteligencia artificial "piensa", "entiende" o "es inteligente", Floridi propone preguntar qué tipo de capacidad de actuar estamos introduciendo en el mundo. Distingue, a partir de esta idea, dos proyectos: por un lado, la inteligencia artificial como agencia artificial; por otro, la agencia política como agencia colectiva transformada por las interacciones digitales.

Floridi distingue entre agencia e inteligencia. Tradicionalmente tendemos a pensar que un agente exitoso debe ser inteligente: actúa bien porque comprende, interpreta, razona o decide. Floridi sostiene que la inteligencia artificial muestra algo distinto: puede haber sistemas que actúan con éxito sin comprender en sentido humano. En este sentido, la verdadera revolución de la inteligencia artificial sería pragmática antes que cognitiva: no consiste en crear mentes artificiales semejantes a las humanas, sino en producir artefactos capaces de realizar tareas de manera eficaz sin requerir comprensión, conciencia, experiencia, sabiduría ni semántica humanas. Esta tesis se expresa con una fórmula muy fuerte: *agere sine intelligere*, actuar sin entender. La inteligencia artificial puede resolver problemas, generar resultados, optimizar procesos o ejecutar acciones moralmente relevantes sin que por ello debamos atribuirle inteligencia en sentido estricto. Esto quiere decir que las máquinas se han vuelto más capaces (no más inteligentes), gracias a que nosotros mismos hemos ido transformando el mundo para hacerlo más habitable para ellas.

A esta especie de rediseño del mundo Floiridi lo llama *enveloping* y lo describe como una recomposición del entorno para volverlo legible, navegable y operable por sistemas digitales. Un ejemplo: no es que los coches autónomos tengan éxito porque posean inteligencia humana; más bien, lo consiguen allí donde las calles, sensores, mapas, señales, datos, normas y entornos estén lo suficientemente digitalizados. Según Floridi, los drones, robots, bots y algoritmos funcionan mejor porque el entorno se ha vuelto cada vez más propicio para las capacidades limitadas de la inteligencia artificial cognitiva. En otras palabras, la inteligencia artificial ha podido surgir en una infosfera, es decir, un ambiente cada vez más informacional, digitalizado y estructurado para agentes artificiales. Por ello, sostiene Floridi que el éxito de la inteligencia artificial depende, por una parte, de la separación entre agencia e inteligencia y, por otra, de la transformación progresiva de nuestros entornos en espacios favorables para ella.

Ahora bien, los agentes artificiales pueden afectar derechos, intereses, oportunidades, recursos y decisiones humanas. Es pertinente, por ello, hablar de una ética de la inteligencia artificial. En este caso, la premisa principal es obvia, aunque, al parecer, no lo es para todo el mundo: la responsabilidad debe seguir recayendo en agentes humanos. Cuando se atribuye el daño al "algoritmo" o al "sistema", se corre el riesgo de desresponsabilizar a quienes diseñaron, implementaron, desplegaron o se beneficiaron de esa agencia artificial. En el ámbito jurídico, Floridi advierte que responsabilizar exclusivamente al agente artificial debilitaría la función disuasoria del derecho penal y desplazaría, sin justificación alguna, la responsabilidad hacia ingenieros, usuarios, empresas o instituciones. Por ello, la pregunta ética importante debería ser más

bien cuánta agencia humana se conserva, se amplía o se delega. Cuando usamos la inteligencia artificial, cedemos parte de nuestro poder de decisión a los artefactos tecnológicos. Floridi piensa, a diferencia de los tecnopesimistas, que ello no es necesariamente malo: delegar puede ser razonable, eficiente e incluso moralmente deseable. El problema surge, sin embargo, cuando la delegación erosiona la autonomía humana.

Ante el riesgo de dicha pérdida de autonomía, Floridi introduce una noción: la de "meta-autonomía": no puede conferirse a la inteligencia artificial la capacidad superior de decidir qué decisiones pueden delegarse, bajo qué condiciones, con qué límites y con qué posibilidad de reversión. La inteligencia artificial no es sólo una herramienta pasiva; es una agencia artificial. En consecuencia, sus efectos deben enmarcarse en categorías como la autonomía, la justicia, la no maleficencia, la beneficencia, la explicabilidad y la responsabilidad. Estos son los cinco principios, según Floridi, de la ética de la inteligencia artificial. Tienen sentido porque apuntan a las preguntas correctas: cómo funciona el sistema y quién responde por su funcionamiento.

Floridi deja claro que la inteligencia artificial no es una simple herramienta sin autonomía funcional. Para definirla, encuentra un término medio: no es una persona capaz de sustituir la inteligencia humana, pero tampoco es un simple martillo. Suena convincente. Sin embargo, creo que Floridi deja algunos cabos sueltos. Me parece que su noción de "agencia" es tan amplia que, si en efecto la entendemos como "interactuar exitosamente con el mundo en vista de un objetivo", muchos sistemas técnicos podrían ser agentes en sentido débil y se correría el riesgo de entender el fenómeno de la inteligencia artificial como un asunto de "intensidad" moderable. Además, la noción de "objetivo" en relación con la de "agencia" es ambigua: ¿el objetivo pertenece al sistema, al diseñador, al usuario, al modelo de negocio o a la institución? Aunque el planteamiento de Floridi, en mi opinión, es en esencia acertado, enciendo una última alerta: la idea de "agentes artificiales", tal como la entiende, podría reforzar —pese a sus propias cautelas— la tentación social de atribuirles demasiada autonomía y descargar en ellos la responsabilidad humana. Esta última preocupación se ha reflejado en varias publicaciones recientes. Me referiré tan sólo a tres ejemplos que ilustran a la perfección los riesgos de la agencia artificial.

En *Genesis: Artificial Intelligence, Hope, and the Human Spirit* (2024), Henry A. Kissinger, Craig Mundie y Eric Schmidt reflexionan sobre el tipo de mundo que ha ido emergiendo desde que hemos permitido

que las máquinas comiencen a producir conocimiento, tomar decisiones, intervenir en la política, transformar la ciencia y alterar la autocomprensión de la humanidad. Este libro reúne a tres autores con perfiles muy distintos: Henry A. Kissinger, figura central de la diplomacia estadounidense del siglo XX; Eric Schmidt, antiguo CEO de Google; y Craig Mundie, exdirectivo de Microsoft. Tras plantear que la inteligencia artificial inaugura una nueva etapa histórica porque separa, de un modo inédito, el conocimiento humano de la comprensión humana, los autores sostienen que la inteligencia artificial introduce resultados que pueden ser eficaces sin ser plenamente inteligibles para nosotros. De ahí el peligro de una nueva forma de autoridad opaca, casi oracular, que los autores vinculan con la posibilidad de una *dark enlightenment*. El libro previene contra la fe ciega en la inteligencia artificial, pero, a la vez, rechaza el temor injustificado hacia ella, de modo que quienes esperaban una crítica devastadora de su uso quizás queden insatisfechos.

El mismo año, 2024, Mariarosaria Taddeo publicó *The Ethics of Artificial Intelligence in Defence* (2024). Su planteamiento aborda un tema preocupante. Se centra en la necesidad de elaborar un marco ético capaz de identificar, analizar y orientar la gobernanza de los usos de la inteligencia artificial en defensa, prestando especial atención a la autonomía, las capacidades de aprendizaje y los niveles de agencia que estas tecnologías habilitan. El libro exige al lector cierta familiaridad tanto con la teoría de la guerra justa como con los debates recientes sobre las tecnologías digitales. Taddeo advierte de los riesgos de la inteligencia artificial cuando se utiliza en labores de logística, análisis de inteligencia, ciberdefensa, ciberguerra, identificación de objetivos, apoyo a la toma de decisiones tácticas e incluso en sistemas de armas autónomos. Este es un caso grave en el que, en efecto, se evidencia que el uso de la inteligencia artificial en defensa combina problemas como la pérdida de control humano, la opacidad, los sesgos y el desplazamiento de responsabilidad, con problemas tradicionales del uso de la fuerza, como la dignidad humana y la posible vulneración de los principios de la guerra justa. Taddeo insiste en que, sin una regulación robusta, la inteligencia artificial puede erosionar los mismos valores que las instituciones de defensa afirman proteger. La ética de la inteligencia artificial es, por lo tanto, una ética de contención: busca hacer gobernable una tecnología cuya lógica tiende a la aceleración, la opacidad y la automatización.

En *Una teoría crítica de la inteligencia artificial* (2025), Daniel Innerarity se pregunta qué ocurre con la democracia cuando una parte creciente de nuestras decisiones queda mediada, condicionada o delegada en sistemas algorítmicos. Al igual que sucede en los otros libros recién mencionados, Innerarity guarda

distancia tanto del entusiasmo tecnosolucionista como del catastrofismo apocalíptico: la inteligencia artificial no salvará automáticamente la democracia ni la destruirá por sí sola; la transformará, la condicionará y nos obligará a repensar sus categorías fundamentales. En el trayecto argumental de su libro Innerarity elabora ideas muy sugerentes. Por ejemplo, sostiene que una crítica adecuada de la inteligencia artificial no puede reducirse ni a una auditoría técnica ni a un código de buenas prácticas. La "algoritmización" obliga a revisar categorías como el sujeto, la acción, la responsabilidad, el conocimiento, el trabajo y la decisión. En este sentido, no se trata de condenar moralmente la tecnología desde fuera, sino de comprender su complejidad para mostrar que sus configuraciones actuales no son inevitables.

Lo que más preocupa a Innerarity es el futuro de la democracia. A ese respecto, observa que la democracia no consiste únicamente en producir buenos resultados, sino en garantizar procedimientos que permitan a los ciudadanos participar en la formación de la voluntad colectiva. Una gobernanza basada en recomendaciones, preferencias implícitas y huellas digitales corre el riesgo de reducir al ciudadano a un consumidor satisfecho o a un generador de datos. Sin embargo, la voluntad popular no equivale a la mera agregación automática de deseos individuales, sino que exige deliberación, conflicto, revisión pública de los fines y autorización colectiva. Precisamente por su complejidad, la inteligencia artificial nos enfrenta a los límites del conocimiento humano y a la necesidad de diseñar formas institucionales de control compatibles con la automatización.

No es posible mantener un entusiasmo acrítico ante un fenómeno como el que tenemos enfrente. La cuestión no se reduce a si la inteligencia artificial es una buena herramienta o no. El debate tampoco puede limitarse a la pregunta de si las máquinas llegarán a ser más capaces que los seres humanos. La verdadera preocupación ha de centrarse, como bien ha dicho Larson, en plantearnos si la inteligencia artificial podrá comprender, conjeturar, interpretar y actuar en contextos abiertos como lo hacemos los seres humanos. Ante esta posibilidad, se ha vuelto indispensable defender nuestra inteligencia y nuestra autonomía. Si hay quienes están creando nuevos agentes artificiales capaces de transformar el mundo, ¿seremos capaces de seguir siendo autores responsables de nuestro propio destino?