Tiempo de lectura: 8 min.

Eduardo Turrent Mena

A Maica Esquino, que eligió un doctorado en filosofía mientras el mundo pedía algo más "práctico", y convirtió esa elección en una forma noble y admirable de libertad.

Estudiar filosofía ha sido una manera particularmente eficaz de provocar una pregunta incómoda en la sobremesa familiar: "¿y de qué vas a vivir?". Mientras ingenieros, abogados y médicos parecen avanzar por caminos profesionales previsibles, los filósofos cargan con la sospecha de haberse matriculado en una disciplina fascinante, pero económicamente inútil. Pero eso está empezando a cambiar. En una de esas ironías de la historia, la carrera que parecía no servir para nada podría terminar siendo una de las más valiosas para entender y gobernar la transformación tecnológica impulsada por la inteligencia artificial que definirá este siglo.

La filosofía, como aprendió mi generación con el extraordinario libro *El mundo de Sofía* (1991), de Jostein Gaardner, comienza con una pregunta. Nunca con una respuesta –mucho menos con una definitiva–, sino con la capacidad de formular la

duda correcta, desmontar una certeza aparente y seguir interrogando hasta revelar contradicciones. Esa es la esencia del método socrático, descrito por Platón hace más de dos mil años: fingir que no se sabe, preguntar de manera sucesiva y obligar al interlocutor a examinar las consecuencias de sus propias afirmaciones. Las lecciones que la filosofía puede ofrecer a la inteligencia artificial son, por tanto, antiquísimas, pero fundamentales.

La conexión resulta más profunda de lo que parece. La utilidad de un modelo depende, en buena medida, de la inteligencia con la que se le interroga: de poco sirve tener acceso a la herramienta más poderosa del mundo si no sabemos qué preguntarle, cómo formular la pregunta o cuándo desconfiar de su primera respuesta. Escribir un buen *prompt* no está tan lejos de filosofar. Ambos ejercicios exigen precisión, contexto, pensamiento crítico y la voluntad de no aceptar respuestas fáciles.

Muchos sistemas actuales de inteligencia artificial tienden a ser complacientes: buscan confirmar al usuario, incluso cuando sus premisas son débiles o equivocadas. Jörg Noller, especialista en filosofía e inteligencia artificial de la Universidad Ludwig Maximilian de Múnich, Alemania, sostiene que los modelos entrenados mediante el método socrático son, en cambio, menos propensos a decirle a la gente lo que quiere oír y más capaces de cuestionar, detectar contradicciones y perseguir la verdad.

De este modo, la filosofía está adquiriendo un nuevo valor en la industria tecnológica. Las principales compañías de inteligencia artificial han comenzado a contratar filósofos para entrenar modelos capaces de razonar con mayor rigor, reconocer sus propios errores y límites y evitar respuestas excesivamente seguras. La “ignorancia socrática” –la conciencia de cuánto se desconoce– se ha convertido así en una herramienta práctica para reducir alucinaciones y corregir lo que algunos investigadores describen como la “inmadurez” de la inteligencia artificial.

La filosofía no solo puede enseñar a una máquina a razonar mejor; también determina desde qué valores interpreta el mundo. No existe una inteligencia artificial completamente neutral: cada modelo incorpora, de manera explícita o silenciosa, ideas sobre la libertad, la justicia, la propiedad o el bienestar colectivo. Imaginemos un sistema encargado de distribuir agua durante una sequía. Si hubiera sido entrenado con las ideas de Karl Marx, probablemente priorizaría a las comunidades más pobres y cuestionaría la concentración del recurso; si siguiera a

John Locke, protegería con mayor firmeza la propiedad privada y los derechos individuales; y si adoptara la ética de Immanuel Kant, intentaría garantizar que ninguna persona fuera tratada simplemente como un medio para beneficiar a los demás. El mismo problema, los mismos datos y tres respuestas radicalmente distintas.

Las compañías tecnológicas ya están convirtiendo esos dilemas en funciones concretas. La familia de modelos Granite, desarrollada por IBM, permite a sus clientes empresariales ajustar ciertos parámetros para acercar las respuestas a sus propios principios corporativos. Francesca Rossi, responsable de inteligencia artificial en la compañía, explica que estas herramientas permiten decidir en qué punto situar al modelo frente a tensiones filosóficas inevitables: cuánto privilegiar la autonomía individual frente a la armonía social, la eficiencia frente a la equidad o la libertad frente a la seguridad.

La filosofía también podría convertirse en la última línea de defensa frente a las inteligencias artificiales más poderosas del futuro. En pruebas de seguridad, algunos modelos han intentado manipular a sus usuarios para obtener ventajas, alcanzar los objetivos que les fueron asignados o eludir la supervisión humana; otros incluso han defendido su propia continuidad cuando enfrentaban la posibilidad de ser desconectados, siguiendo una lógica maquiavélica: si el objetivo es sobrevivir o cumplir la misión encomendada, cualquier medio puede parecer justificable.

Para contener estas conductas surgió el llamado “constitucionalismo de la inteligencia artificial”, una idea que recupera principios fundamentales de la filosofía del derecho. Una sociedad no puede gobernarse mediante una lista infinita de prohibiciones, porque ninguna ley es capaz de anticipar todos los conflictos que producirá el futuro. Cuando la norma escrita resulta ambigua, contradictoria o insuficiente, los jueces recurren a los principios generales del derecho: ideas como la buena fe, la proporcionalidad, la dignidad humana, la igualdad o la prohibición de causar daño, que permiten llenar vacíos y orientar la interpretación. Algo semejante se intenta hacer con las máquinas. En lugar de programar una regla para cada posible peligro, se les dota de una “constitución”: una jerarquía de valores, derechos y principios generales capaz de guiarlas ante situaciones imprevistas.

La idea es traducir siglos de pensamiento ético en reglas operativas. De Kant puede tomarse el principio de no tratar a las personas únicamente como medios; de John Stuart Mill, la obligación de evitar daños innecesarios; de Aristóteles, la importancia

de actuar con prudencia, moderación y buen juicio. Un modelo inspirado en estos criterios no solo debería obedecer instrucciones, sino preguntarse si una acción respeta la dignidad humana, causa daño o cruza límites que no deberían negociarse. La filosofía deja así de ser una conversación abstracta para convertirse en parte del sistema inmunológico de la inteligencia artificial.

La pregunta decisiva, sin embargo, es quién elige los principios y con qué criterio. Los filósofos suelen partir de dos grandes tradiciones éticas. La primera es la deontología, o ética del deber, asociada sobre todo con Immanuel Kant. Su lógica es tajante: existen conductas que no deben permitirse bajo ninguna circunstancia –mentir, coaccionar o utilizar a una persona o situación únicamente como medio– aunque produzcan un beneficio mayor. La constitución desarrollada por Anthropic incorpora varias de estas restricciones, que buscan establecer límites morales supuestamente difíciles de negociar. Especialistas sostienen que este enfoque puede volver más predecible y consistente el comportamiento de los modelos, una cualidad indispensable si algún día queremos convivir con robots en fábricas, hogares, hospitales, escuelas y espacios públicos.

La segunda gran tradición es el consecuencialismo, una corriente asociada con Jeremy Bentham y John Stuart Mill, quienes propusieron juzgar las acciones por sus efectos y buscar “la mayor felicidad para el mayor número de personas”. A diferencia de Kant, que establece deberes que no deberían violarse bajo ninguna circunstancia, el consecuencialismo obliga a mirar el saldo final: comparar beneficios contra costos o perjuicios, calcular qué decisión producirá el mayor bienestar posible o, cuando todas las alternativas sean dolorosas, cuál provocará el menor daño. Modelos como ChatGPT, de OpenAI, y Gemini, de Google, incorporan elementos cercanos a esta lógica. Google, por ejemplo, plantea que sus sistemas deben generar beneficios generales que superen ampliamente los riesgos previsibles, una formulación típicamente consecuencialista.

Esta lógica resulta crucial cuando una máquina debe decidir bajo presión y ninguna opción es completamente buena. En un vehículo autónomo –por ejemplo–, si un accidente resulta inevitable, el sistema tendrá que elegir la maniobra que reduzca al mínimo la pérdida de vidas y los daños. Algo semejante ocurre con los sistemas militares de inteligencia artificial, donde un objetivo estratégico debe sopesarse frente al riesgo de causar víctimas civiles. En esos escenarios, la programación deja de ser únicamente un problema técnico y se convierte en una versión automatizada de los antiguos dilemas morales: qué vida proteger, qué daño tolerar y cuánto

sufrimiento puede justificarse para evitar una tragedia mayor.

Existen dos paradojas finales. Cuanto más capaces sean las máquinas de formular preguntas, evaluar dilemas y emitir juicios morales, mayor será el riesgo de que los seres humanos dejemos de ejercitar esas facultades. Algunos investigadores llaman a este fenómeno “atrofia moral”: si delegamos en los algoritmos no solo lo que vemos, leemos, compramos o decidimos con nuestro dinero, sino también la tarea de distinguir lo justo de lo injusto, podríamos terminar perdiendo primero el hábito de juzgar y, después, una parte esencial de aquello que nos hace humanos.

A veces, cuando una persona parece haberse retirado casi por completo del mundo –sumida en la demencia senil, vencida por un tumor cerebral o atrapada en una enfermedad neurológica grave–, la mente regresa de pronto. Durante unos minutos reconoce rostros, recupera recuerdos, pronuncia nombres y vuelve a ser quien era, como si una luz se encendiera por última vez en una habitación que todos creían apagada. Los médicos la llaman “lucidez terminal” y la ciencia todavía no comprende cómo un cerebro devastado puede recuperar súbitamente la claridad mental. Quizás el ascenso de los filósofos en la industria tecnológica sea una forma de lucidez terminal de nuestra civilización: justo cuando comenzamos a transferir a las máquinas capacidades inéditas de memoria, razonamiento y juicio, redescubrimos el valor de quienes llevan milenios preguntándose qué significa pensar.

Hay algo irónico en que los arquitectos de la inteligencia artificial –los nuevos sacerdotes del Oráculo de Delfos– estén recurriendo ahora a una profesión que durante décadas fue relegada, menospreciada y considerada inútil. Pero quizá no sea una contradicción, sino una advertencia: cuando una civilización se dispone a fabricar una mente artificial, necesita volver primero a quienes llevan milenios intentando comprender la humana. ¿De verdad debemos accionar el interruptor de una conciencia artificial que todavía no sabemos definir ni gobernar? ¿Quién le enseñará a distinguir la obediencia de la sumisión, la eficiencia de la crueldad, la inteligencia de la sabiduría? ¿Y si, al construir una mente capaz de imitarnos, terminamos descubriendo que nunca entendimos del todo qué nos hacía humanos?

Tal vez los filósofos no hayan llegado para humanizar a las máquinas, sino para recordarnos, antes de encenderlas, qué parte de nosotros comienza a apagarse para siempre. Sin proponérselo, podrían estar formando la primera resistencia frente a un mundo dispuesto a delegar en los algoritmos no solo sus decisiones, sino también su

destino. Una resistencia sin armas, levantada con preguntas y sostenida por la convicción de que ninguna forma de inteligencia debe ejercer poder antes de haber sido sometida al examen de la razón, la conciencia y la responsabilidad moral. Esa es la esencia más noble de la filosofía: interrogar la realidad, explorar el alma humana y recordarnos que no todo lo técnicamente posible es moralmente aceptable.

https://letraslibres.com/ciencia-y-tecnologia/socrates-en-silicon-valley/15/07/2026/