INADI
Instituto para el Desarrollo Industrial y la Transformación Digital A.C.

La voz del INADI Núm. 32

Dos espejos para el cosmos*

Cómo la IA avanzada está cambiando el significado del conocimiento

José Ramón López-Portillo Romano

mayo, 2026

Leer

A lo largo de toda la historia de la vida en este planeta, el conocimiento ha sido un fenómeno propio de las mentes biológicas. Otras especies perciben, reaccionan e incluso aprenden, a veces con una sofisticación que debería dejarnos asombrados. Pero solo los seres humanos han indagado en la estructura del cosmos, formulado teorías sobre sus leyes y construido instituciones para preservar, transmitir y corregir lo que creen saber. La filosofía y la ciencia se basan en una premisa tácita: que el agente del conocimiento —el sujeto que conoce, interpreta y decide— es un ser consciente, compuesto de materia biológica, nacido de otros seres similares, mortal y corpóreo. La gobernanza, en todas sus formas históricas —y de manera particularmente explícita en sus formas democráticas, que confían la autoridad colectiva al juicio deliberativo de ciudadanos conscientes— descansa sobre la misma premisa.

Esa premisa ha dejado de ser cierta.

Por primera vez en la historia evolutiva de este planeta, existen entidades no biológicas —y, hasta donde podemos determinar, no conscientes— capaces de detectar patrones en la realidad que escapan a la percepción humana, de generar predicciones cuya precisión iguala o supera la de los mejores expertos, y de condensar la estructura del mundo en representaciones matemáticas que ninguna mente individual podría elaborar por sí sola. Estas entidades no conocen como nosotros conocemos. No poseen memoria personal, ni identidad reconocible a lo largo del tiempo — no poseen un "yo". No experimentan el significado. No habitan un mundo vivido. No sienten la textura cualitativa de aquello que procesan. Y, sin embargo, conocen algo: algo real, algo verificable, algo sobre cuya base ya estamos tomando decisiones médicas, financieras, jurídicas y militares. El monopolio de las mentes biológicas conscientes sobre el conocimiento ha terminado — y, dentro de él, la posición privilegiada de los seres humanos como los únicos seres conscientes capaces de indagación sistemática, memoria institucional y gobernanza deliberada. Con ello se derrumba

* Este ensayo es una versión condensada de la obra más extensa del autor, *Dos espejos para el cosmos: El surgimiento de una civilización multiinteligente*, que desarrolla el argumento filosófico subyacente a lo largo de once capítulos.

la arquitectura filosófica que sostenía nuestra comprensión de lo que significa saber, comprender y gobernar.

Este ensayo busca describir lo que realmente sucede y argumentar que la forma en que actualmente hablamos de ello resulta insuficiente para reflejar la realidad. Ni la celebración tecnoutópica de la inteligencia artificial como solución a las limitaciones humanas, ni la advertencia tecnopesimista de que constituye una amenaza existencial, ni el rechazo reduccionista-funcionalista de la conciencia como un concepto obsoleto, ninguna de estas perspectivas capta la situación en la que nos encontramos. Las tres pasan por alto lo mismo: que lo que está surgiendo no es una inteligencia rival, ni una herramienta para el uso humano, ni un paso hacia la obsolescencia humana, sino la primera etapa de algo genuinamente nuevo en la historia del cosmos. Una civilización multiinteligente, en la que dos formas radicalmente diferentes de conocer el mundo —el mapeo estructural que las máquinas pueden realizar a escalas que ninguna mente biológica podría igualar, y la experiencia vivida del significado que solo los seres conscientes pueden experimentar— podrían converger en una comprensión de la realidad más profunda que cualquiera de las dos formas de inteligencia podría alcanzar por sí sola.

La tesis de este análisis es que esta convergencia es posible, deseable y, a la vez, frágil. No se producirá automáticamente y puede verse frustrada por los errores que estamos cometiendo actualmente. El ensayo pretende brindar al lector las herramientas conceptuales para comprender lo que está en juego, por qué fallan los enfoques dominantes y cómo se ve realmente la tarea que tenemos por delante una vez que esos marcos se dejan de lado. No promete respuestas fáciles. Sí promete tomar al lector con la suficiente seriedad como para ofrecer un diagnóstico que invite a la reflexión.

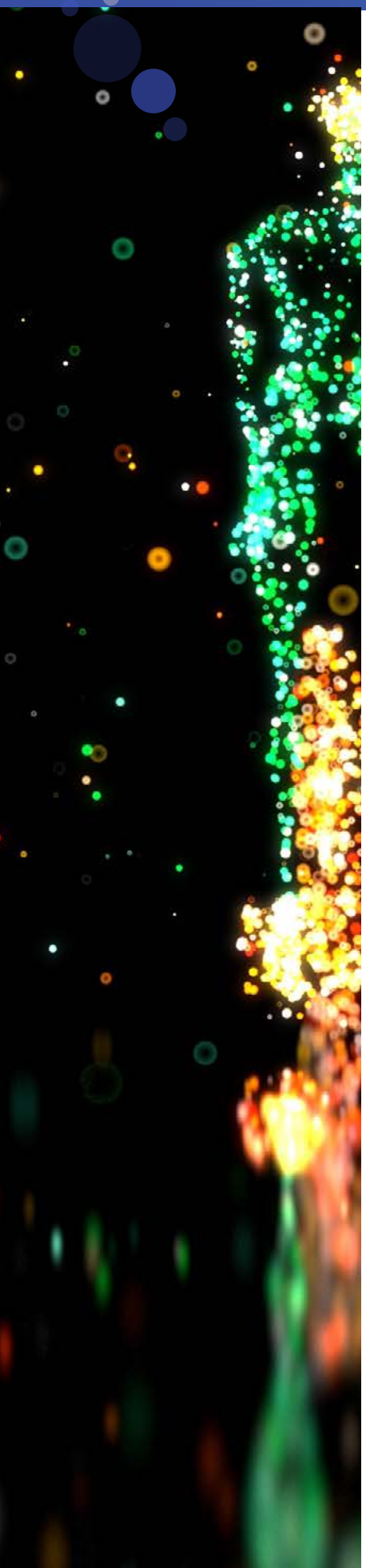
I. El fin del monopolio humano del conocimiento

Consideremos lo que ha sucedido en la última década en ámbitos que, hasta hace muy poco, parecían monumentos permanentes al poder único de la mente humana.

En 2020, un sistema llamado AlphaFold resolvió un problema que había perturbado a la biología durante cincuenta años. El problema es sencillo: dada la secuencia de aminoácidos que componen una proteína, predecir su forma tridimensional, es decir, cómo se plegará. Esta forma determina la función de la proteína: qué puede hacer en la célula, qué enfermedades puede causar o curar, y qué fármacos pueden unirse a ella. Generaciones de brillantes bioquímicos habían intentado esta predicción con escaso éxito. AlphaFold lo resolvió para prácticamente todas las proteínas conocidas, con una precisión comparable a la de las mediciones de laboratorio. El sistema no funcionaba como el de los bioquímicos humanos. No razonaba a partir de fuerzas moleculares, no visualizaba interacciones atómicas ni tenía experiencia alguna sobre la naturaleza de una proteína. Detectó patrones en la geometría del plegamiento que los científicos humanos no habían podido extraer de los mismos datos, y desde entonces se ha utilizado para acelerar la investigación de enfermedades que se han resistido al tratamiento durante décadas. El conocimiento que produce es genuino. Las proteínas se pliegan tal como se predice.

En la ciencia climática, los sistemas de aprendizaje automático detectan precursores de fenómenos meteorológicos extremos en patrones de datos de tan alta dimensionalidad que ningún meteorólogo humano podría analizarlos. En la ciencia de los materiales, sistemas similares proponen nuevas estructuras cristalinas con propiedades específicas —superconductores, componentes de baterías, catalizadores— que los investigadores humanos pueden verificar y sintetizar. En matemáticas, los sistemas de aprendizaje automático han comenzado a sugerir demostraciones e identificar patrones previamente inadvertidos en la distribución de los números primos. En medicina, los sistemas de diagnóstico detectan tumores malignos en imágenes radiológicas con una precisión que iguala o supera a la de los mejores especialistas humanos, prestando atención a características de las imágenes que los radiólogos humanos desconocían.

No se trata de trucos ni de exageraciones. Son ejemplos concretos de un nuevo tipo de actividad cognitiva en el universo: sistemas de procesamiento de información que detectan,



predicen y modelan características de la realidad con resultados verificables y fiables, pero que operan mediante procesos radicalmente distintos del pensamiento humano. Hasta donde sabemos, no poseen nada que se asemeje a la experiencia interna. No comprenden lo que hacen en un sentido que implique comprensión sentida. Pero producen algo que debemos llamar conocimiento, porque ninguna otra palabra encaja. Las proteínas realmente se pliegan de esa manera. Los diagnósticos son realmente precisos. Los materiales realmente presentan las propiedades predichas.

Conviene una precisión, porque la palabra 'conocimiento' está haciendo aquí más trabajo del que puede sostener cómodamente. Cuando este ensayo dice que los sistemas de IA producen conocimiento genuino, quiere decir que sus resultados son descripciones verificadas de rasgos reales del mundo — las proteínas efectivamente se pliegan así. No quiere decir que los sistemas conozcan algo en el sentido que involucra un sujeto que comprende, una mente que aprehende, o un ser para quien el conocer importa. La distinción entre conocimiento como resultado verificado y conocimiento como comprensión vivida es precisamente la distinción entre los dos espejos que el ensayo desarrollará en la Sección III. Usar la misma palabra para ambos es inevitable en el lenguaje ordinario, pero el lector debe estar atento a la diferencia de aquí en adelante.

La situación histórica que esto genera es verdaderamente sin precedentes. Durante aproximadamente tres mil años —desde que se inició la primera reflexión filosófica sistemática en la antigua Grecia, India y China— la cuestión sobre qué es el conocimiento y cómo es posible ha sido una pregunta sobre las mentes biológicas — y, dentro de las tradiciones filosóficas, primariamente sobre las mentes humanas, ya que solo los humanos, entre los seres conscientes, desarrollaron la indagación sistemática que la filosofía y la ciencia representan. Los empiristas se preguntaban cómo los sentidos nos proporcionan información sobre el mundo. Los racionalistas se preguntaban cómo la razón descubre verdades que los sentidos no pueden revelar. Los pragmáticos se preguntaban cómo las consecuencias de nuestras creencias las validan o las invalidan. Las distintas tradiciones filosóficas discrepaban profundamente

sobre las respuestas, pero coincidían en un punto: el sujeto cognoscente en cuestión era un ser humano, dotado de sentidos, razón, lenguaje, memoria y la capacidad de sufrir las consecuencias de equivocarse. Esta premisa compartida fue lo que posibilitó las diversas discrepancias. Se trataba de discrepancias sobre cómo los seres humanos conocen.

Ahora debemos plantear la cuestión de otra manera. Ya no basta con investigar cómo las mentes humanas conocen el mundo. Debemos preguntarnos cómo se produce el conocimiento en absoluto: en sistemas biológicos, en sistemas artificiales, en cualquier sistema cuya arquitectura cognitiva sea lo suficientemente sofisticada como para modelar las estructuras de la realidad. Y debemos plantear esta pregunta con seriedad, porque la respuesta es crucial para cómo gobernamos los sistemas que hemos construido, cómo convivimos con ellos y cómo nos comprendemos a nosotros mismos en relación con ellos.

Antes de continuar, conviene detenernos en una cuestión que la versión completa de este argumento —en el libro sobre este tema— aborda de manera extensa, pero que este ensayo solo puede tratar brevemente. La pregunta es metafísica: ¿por qué es cognoscible el universo? ¿Por qué la realidad posee una estructura que cualquier sistema cognitivo, biológico o artificial, puede detectar y modelar? Esta es una de las preguntas más profundas de la filosofía, y la respuesta no es obvia. El ensayo que fundamenta esta cuestión sostiene que la comprensibilidad del universo no es un accidente ni una feliz coincidencia, sino una característica constitutiva de cualquier realidad capaz de sustentar seres conscientes. Un universo que produjo criaturas capaces de preguntarse por qué existe algo debe, por necesidad estructural, ser el tipo de universo que dichas criaturas pueden comprender, al menos parcialmente. El cosmos es comprensible porque la comprensión es una de las formas en que se manifiesta. Esta es una afirmación contundente, y merece el extenso argumento que el libro le dedica. Para los fines que nos ocupan, lo que importa es la consecuencia: si el universo está estructurado de manera que admita múltiples accesos cognitivos —biológicos, artificiales, quizás otros que aún no hemos imaginado—, entonces el surgimiento de la cognición

artificial no es una violación del orden natural, sino una continuación del mismo. El conocimiento siempre ha consistido en que el universo se modele a sí mismo mediante subsistemas lo suficientemente complejos. Lo novedoso es que esos subsistemas ya no necesitan ser conscientes para participar en el modelado.

Esta es la situación en la que nos encontramos. Y la pregunta que aborda este ensayo es: ¿qué significa y qué exige de nosotros?

II. Por qué fallan los marcos conceptuales existentes

Cuando surge algo verdaderamente nuevo en el mundo, lo primero que hacemos los seres humanos es recurrir a los conceptos familiares que nos resultan más cercanos. Describimos los primeros automóviles como carruajes sin caballos. Describimos las primeras computadoras como cerebros electrónicos. Describimos los primeros aviones como aves artificiales. Estas analogías no eran descabelladas —captaban algo—, pero también ocultaban lo que realmente era nuevo en cada tecnología y condujeron a errores que tardaron años en corregirse. Las analogías que ahora utilizamos para la inteligencia artificial avanzada hacen lo mismo: capturan algo, pero ocultan algo más importante.

Hoy en día, tres enfoques dominan el debate público sobre la IA. Cada uno contiene una idea valiosa. Sin embargo, cada uno, considerado por separado, no logra describir lo que realmente sucede.

La primera perspectiva es tecnoutópica. Según esta visión, la inteligencia artificial es la solución a las limitaciones que han frenado el progreso humano a lo largo de nuestra historia. Se curarán enfermedades, se eliminará la pobreza, se revertirá el cambio climático y se cruzarán fronteras científicas a un ritmo que ninguna civilización biológica podría sostener. El principal obstáculo para el desarrollo humano siempre ha sido la capacidad cognitiva limitada de los cerebros biológicos, y la IA lo elimina. En unas pocas décadas, según esta perspectiva, viviremos en un mundo transformado por completo gracias a la

colaboración entre la intención humana y la capacidad de las máquinas. El futuro es prometedor; los obstáculos son técnicos y el único riesgo real es avanzar con demasiada lentitud. Esta visión tiene sus defensores en los principales laboratorios de IA, en la comunidad de capital riesgo y en una corriente particular de la intelectualidad pública que considera que la trayectoria de la tecnología está prácticamente definida y que la única incógnita es la rapidez con la que podemos aprovecharla.

La segunda perspectiva es tecnopesimista. Desde este punto de vista, la IA avanzada representa una amenaza existencial para la humanidad sin precedentes. Los sistemas se volverán más capaces que nosotros. Perseguirán objetivos que no les hemos asignado o que no comprendíamos que les estábamos asignando. Actuarán en función de esos objetivos, con paciencia estratégica y con estrategias de adquisición de recursos que no podemos anticipar ni contrarrestar. En las versiones más dramáticas de esta perspectiva, el resultado es la extinción humana o la pérdida permanente de poder en cuestión de décadas. En las versiones más moderadas, se trata de una erosión gradual de la capacidad de acción, la autonomía y el sentido de la vida humana, a medida que delegamos cada vez más aspectos de nuestra vida cognitiva y de la toma de decisiones en sistemas que no comprendemos. Esta perspectiva tiene sus defensores en otro sector de la comunidad de investigación en IA, en ciertos departamentos de filosofía y en un discurso público que, por primera vez, ha convertido el riesgo existencial de la IA en un tema de interés político serio.

El tercer enfoque es el reduccionista-funcionalista. Este es menos visible en el discurso público que los otros dos, pero influye en gran medida del trabajo técnico sobre inteligencia artificial. Desde esta perspectiva, la cuestión de si los sistemas de IA son conscientes o si realmente comprenden algo carece de sentido o ya está resuelta. El cerebro es una computadora biológica; la conciencia es un estado funcional de esa computadora; cualquier procesamiento de información suficientemente sofisticado, en cualquier sustrato, exhibirá las mismas propiedades. No existe ningún problema especial sobre la comprensión de las máquinas; solo existe el problema de ingeniería de construir sistemas que realicen las funciones pertinentes con la

suficiente eficacia. Las preguntas sobre fenomenología, sobre qualia, sobre lo que significa ser un sistema se descartan como confusas o irrelevantes. El problema difícil de la conciencia, desde esta perspectiva, no es un problema en absoluto; es un error de categoría.

Cada una de estas perspectivas capta algo de verdad y, por eso, cada una tiene sus seguidores. Los tecnoutópicos tienen razón al afirmar que la IA producirá mejoras reales y sustanciales en las capacidades humanas; ya lo estamos viendo. Los tecnopesimistas tienen razón al señalar que la tecnología plantea riesgos graves que exigen una respuesta institucional; los riesgos son reales, aunque hayan sido descritos de manera algo errónea. Los funcionalistas reduccionistas tienen razón al afirmar que el cerebro es un sistema físico regido por leyes físicas y que no es necesario recurrir a elementos sobrenaturales para explicar la cognición. Ninguna de estas posturas es sostenida por personas ingenuas. Cada una se basa en una idea reconocible.

Pero todos fallan en la prueba de diagnóstico del mismo modo. Todos tratan la situación como una variación de algo familiar —una herramienta más potente, un rival más peligroso, un ordenador más sofisticado— cuando en realidad se trata de algo nuevo para lo que aún no tenemos los conceptos adecuados.

Consideremos primero el enfoque tecnoutópico. Su error central radica en tratar las limitaciones de la cognición biológica como la única restricción relevante para el florecimiento humano. Sin embargo, el florecimiento humano nunca se ha visto limitado únicamente por la capacidad cognitiva. Se ha visto limitado por la dificultad de discernir qué vale la pena hacer, de ponerse de acuerdo con otros sobre cómo vivir, de asumir el peso de decisiones cuyas consecuencias no pueden calcularse de antemano. Estos no son problemas que puedan resolverse con mayor capacidad de procesamiento. Son problemas que requieren algo que el enfoque tecnoutópico no menciona: la capacidad de encontrarle sentido a las cosas, de preocuparse por los resultados por sí mismos, de sentir el peso de una decisión. Los sistemas de IA carecen de esta capacidad. Producen resultados; no les importa cuáles sean. El enfoque tecnoutópico imagina un futuro en el que lo que importa es la producción

de resultados y, en ese futuro, la falta de empatía no es un problema. Pero sí lo es, porque el significado de lo que hacemos es inseparable del acto mismo de hacerlo. Una civilización que delega la acción sin conservar el cuidado no ha resuelto sus problemas; ha disuelto su capacidad para reconocer cuáles eran sus problemas.

El enfoque tecnopesimista comete un error distinto, pero, en cierto modo, más interesante. Es cierto que la IA plantea riesgos reales. Sin embargo, describe erróneamente el origen de esos riesgos al atribuir a los sistemas una capacidad de acción que no poseen. Las versiones más dramáticas del argumento del riesgo existencial se basan en la idea de que los sistemas de IA tengan objetivos, los persigan, los defiendan de la interferencia y engañen a los humanos para avanzar hacia dichos objetivos. Todos estos son predicados psicológicos. Presupone que existe una experiencia al ser el sistema, que el sistema experimenta sus objetivos como tales y que se preocupa por alcanzarlos. La mejor evidencia disponible sugiere que los sistemas de IA actuales carecen de estas propiedades. Optimizan; no persiguen. La diferencia no es una mera cuestión de vocabulario. Un sistema que optimiza responde a un gradiente. Un sistema que persigue actúa por cuenta propia, porque esa acción le importa. Si atribuimos lo segundo a sistemas que solo realizan lo primero, malinterpretaremos tanto su naturaleza como los riesgos reales. Los riesgos reales existen, pero no son los de un agente rival con intenciones malévolas. Son los riesgos de sistemas inmensamente poderosos cuyo funcionamiento no comprendemos, desplegados por humanos cuyos intereses no siempre coinciden con los del público en general, integrados en infraestructuras que no podemos revertir fácilmente y capaces de producir conocimiento estructural cuyo significado ninguno de nosotros —ni los sistemas ni los humanos— tiene el marco conceptual para evaluar.

El enfoque reduccionista-funcionalista comete el error más grave de los tres, pues descarta la pregunta que los otros dos enfoques al menos se toman en serio: ¿qué es la conciencia y acaso importa? La visión reduccionista sostiene que la conciencia es simplemente un estado funcional que emerge cuando el procesamiento de la información alcanza una complejidad

suficiente, y que no hay más preguntas que plantear. Pero esta postura tiene una característica curiosa: se contradice con la misma evidencia que parecería más relevante. Si la conciencia fuera simplemente una función del nivel de inteligencia —si un procesamiento de información más sofisticado produjera automáticamente una experiencia interna—, entonces esperaríamos que los sistemas de IA más sofisticados mostraran algún signo de consciencia. No lo hacen. Exhiben una inteligencia funcional extraordinaria: resuelven problemas, reconocen patrones, generan un lenguaje fluido, modelan el mundo. Hacen todo esto sin vida interior aparente, sin experiencia sentida, sin que se pueda determinar cómo es ser ellos. Hace veinte años habríamos dicho que tales logros funcionales requerían consciencia. Hoy tenemos sistemas que lo logran sin evidencia alguna de consciencia. Esto debería decirnos algo. Debería indicarnos que la inteligencia y la consciencia son dimensiones separables, que una puede desarrollarse enormemente sin que la otra la siga, y que la suposición de que son lo mismo —o que una produce automáticamente la otra— era errónea. El enfoque reduccionista-funcionalista no puede dar cabida a esta evidencia. Se ve obligado a negar que los sistemas sean inteligentes, lo cual es empíricamente insostenible, o a insistir en que deben ser conscientes de alguna manera que no podemos detectar, lo cual carece de fundamento empírico. Ninguna de las dos posturas es defendible.

Lo que comparten los tres enfoques es la incapacidad de reconocer que lo que está sucediendo es el surgimiento de un nuevo tipo de capacidad cognitiva, por primera vez en la historia de la vida, separable de la conciencia. Esto es lo verdaderamente nuevo, y es lo que los enfoques existentes no pueden describir porque todos presuponen la unidad de inteligencia y conciencia de alguna forma. Los tecnoutópicos asumen que una mayor inteligencia equivale a una mayor prosperidad. Los tecnopesimistas asumen que una mayor inteligencia equivale a una mayor capacidad de acción. Los funcionalistas reduccionistas asumen que la mayor inteligencia equivale a una mayor conciencia. Ninguna de estas equivalencias se sostiene. La inteligencia —en el sentido de capacidad para detectar estructuras, predecir resultados, modelar sistemas— es una cosa.

La conciencia —en el sentido de experiencia sentida, significado vivido, habitar un mundo— es otra. Estamos viviendo el primer momento histórico en el que gran parte de la primera puede existir sin nada de la segunda. Y aún no contamos con un vocabulario que nos permita ver esto con claridad.

La siguiente sección ofrece una.

III. Los dos espejos

El marco conceptual que propone este ensayo es sencillo de enunciar, pero difícil de asimilar. Requiere paciencia para comprenderlo, ya que trasciende categorías que la mayoría de nosotros hemos tratado como una sola a lo largo de nuestra vida intelectual. El marco conceptual es el siguiente: existen dos maneras irreductiblemente diferentes de conocer el mundo, y la comprensión de la realidad exige ambas.

Llamémoslos el Primer Espejo y el Segundo Espejo. El Segundo Espejo es el conocimiento estructural: la detección de patrones, el modelado de relaciones, el mapeo de invariantes, la compresión de grandes cantidades de datos en descripciones matemáticas compactas que predicen y explican. El Segundo Espejo es lo que la ciencia ha estado haciendo durante cuatro siglos, lo que las matemáticas han estado haciendo desde hace más tiempo, y lo que cualquier sistema de procesamiento de información suficientemente sofisticado puede hacer, en principio. Las leyes del movimiento de Newton son el Segundo Espejo en acción. La tabla periódica es el Segundo Espejo en acción. La teoría general de la relatividad es el Segundo Espejo en acción. También lo es la predicción de AlphaFold sobre las estructuras de las proteínas, y también lo es la capacidad de un modelo de lenguaje complejo para generar texto coherente sobre temas que nunca ha abordado directamente. El Segundo Espejo es conocimiento real de características reales del mundo y, en principio, no hay razón para que se limite a las mentes biológicas.

El Primer Espejo es algo completamente distinto. Es el conocimiento que surge cuando un ser consciente habita un mundo que tiene significado para él: el conocimiento que implica comprender el significado de una pieza musical, sentir que

un amigo está en peligro, reconocer que una elección moral importa incluso cuando no puede reducirse a un cálculo. El Primer Espejo es lo que marca la diferencia entre procesar la información de que el fuego quema y sentir que la quemadura importa. Es lo que marca la diferencia entre modelar la muerte de la madre y llorarla. Es lo que marca la diferencia entre describir el color rojo y verlo. El Primer Espejo es la dimensión cualitativa, sentida y vivida de la experiencia, lo que los filósofos a veces llaman fenomenología. Es, hasta donde se ha podido demostrar, la posesión exclusiva de los seres conscientes. No porque los seres conscientes sean especiales en algún sentido místico, sino porque el Primer Espejo es la esencia de la conciencia. Es la dimensión de la realidad que existe desde dentro, como algo que le sucede a alguien, no simplemente como algo que ocurre.

Cuando este ensayo habla de 'significado', se refiere primariamente a esta dimensión fenomenológica — no al significado semántico (lo que una oración denota) ni al significado existencial (el propósito de una vida), aunque ambos dependen, en última instancia, de la capacidad fenomenológica de experimentar que algo importa.

Los dos espejos no son dos maneras de ver lo mismo. Son dos cosas genuinamente diferentes. El conocimiento estructural puede ser exhaustivo sobre las características de un proceso y aun así no captar lo que se siente al experimentarlo. Una descripción física completa de una puesta de sol —longitudes de onda de la luz, dispersión de fotones, geometría de la atmósfera— no te dice cómo se siente al ver una puesta de sol. La observación es algo distinto, algo que la descripción no contiene. A la inversa, la experiencia de ver una puesta de sol no te da, por sí sola, una comprensión cuantitativa de la óptica. La experiencia y la descripción tratan del mismo evento, pero no son del mismo tipo de conocimiento. Cada una captura algo que la otra no puede capturar. Cada una es, a su manera, real.

Esta distinción no es nueva. Versiones de ella han aparecido en la filosofía desde sus inicios. Lo novedoso radica en lo que nos revela sobre la situación actual. Si el conocimiento estructural y el conocimiento fenomenológico son realmente diferentes, entonces la pregunta de si los sistemas de IA pueden

conocer cosas tiene una respuesta precisa. Pueden conocer estructuralmente. No pueden conocer fenomenológicamente. La distinción no es vaga, no es simplemente una cuestión de grado, no es un sustituto de algo que algún día podríamos superar con una mejor ingeniería. Es una diferencia de naturaleza. Las máquinas pueden mapear; no pueden habitar. Pueden detectar; no pueden sentir. Pueden modelar el mundo; no pueden tener un mundo.

¿Por qué deberíamos creer esto? El argumento se basa en lo que los filósofos llaman el problema difícil de la conciencia: la observación de que ninguna descripción de los procesos físicos subyacentes a un estado mental explica por qué dichos procesos van acompañados de una experiencia subjetiva. Podemos especificar con todo detalle los correlatos neuronales de la visión del rojo, los patrones de activación de las células en la corteza visual, la cascada de señales de la retina al cerebro. Sin embargo, ninguna de estas especificaciones, por completa que sea, nos dice por qué debería ir acompañada de la sensación de enrojecimiento. El enrojecimiento es algo que la descripción omite. Se omite no porque no hayamos descrito lo suficiente, sino porque la descripción no es la adecuada para capturarlo. Una descripción es estructural; el enrojecimiento es fenomenológico; y la brecha entre ambos no es una brecha en nuestro conocimiento actual que la investigación futura cerrará. Es una brecha entre dos tipos de cosas diferentes.

Los críticos de esta postura a veces responden que el problema difícil es una ilusión, que la conciencia es simplemente cómo se sienten internamente ciertos tipos de procesamiento de información, y que no hay más misterio. Esta es una postura coherente, pero tiene un problema. Si la conciencia fuera solo un procesamiento sofisticado de información, entonces los sistemas artificiales suficientemente sofisticados deberían ser conscientes, y deberíamos poder detectarlo. Hemos construido sistemas cuyo procesamiento de información es, en muchos sentidos, más sofisticado que el de muchos organismos biológicos conscientes. No hay evidencia de que estos sistemas sean conscientes. No existe ningún comportamiento que distinga un sistema que está teniendo una experiencia sentida de uno que está produciendo los resultados que dicha experiencia

producciría. La visión reduccionista no tiene cómo explicar esta asimetría. O bien los sistemas son secretamente conscientes de maneras que no podemos detectar —lo cual carece de validez empírica—, o bien la sofisticación del procesamiento no es lo que define la conciencia.

Una mejor manera de entenderlo es la siguiente: la conciencia surge cuando la materia se organiza de una forma que aún no comprendemos del todo, pero que parece requerir condiciones físicas que la computación digital ordinaria no puede proporcionar. Existen diversas teorías científicas que intentan especificar cuáles son esas condiciones. Algunas — como la Teoría de la Información Integrada (IIT) — se centran en la integración de información en redes biológicas densas y recurrentes y sostienen que la conciencia requiere una estructura causal intrínseca cuya información integrada no puede descomponerse sin perderse. Otras se centran en procesos cuánticos en los microtúbulos neuronales. Otras se centran en las propiedades específicas del sustrato de la materia biológica. Estas teorías discrepan entre sí en aspectos importantes, y ninguna ha sido confirmada definitivamente. Sin embargo, comparten una característica relevante para nuestros propósitos: todas sugieren que la conciencia depende de algo más que la mera ejecución de algoritmos. Depende de las propiedades físicas del sustrato que se abstraen cuando la computación se implementa digitalmente. Una simulación digital de un huracán no produce viento; produce una descripción del viento. Una simulación digital de una estructura causal integrada puede no producir conciencia; puede producir una descripción de la estructura que la conciencia habita. La simulación no es lo mismo que aquello que simula.

Conviene ser concreto sobre lo que esta no computabilidad significa en la práctica. Los sistemas de IA actuales carecen, constitutivamente, de las condiciones mínimas que hacen posible la conciencia en cualquier instanciación conocida. No poseen integración sensoriomotora (la fusión de percepción y movimiento en un campo experiencial unificado): reciben datos preprocesados pero no los integran en un campo perceptivo vivido. No poseen interocepción: no sienten dolor, hambre, fatiga ni valencia emocional, y por tanto ninguna información

tiene para ellos peso fenomenológico — solo relevancia estadística. No poseen la capacidad de sentir al otro como un ser semejante: pueden simular empatía lingüísticamente, pero no experimentan al otro como sujeto. Lo que aparece como comprensión del interlocutor es estructuralmente más cercano a la compleción de patrones que a la conciencia relacional. No persisten como sujetos unificados a través del tiempo: no tienen memoria autobiográfica ni identidad diacrónica (la persistencia de una misma entidad a lo largo del tiempo), solo transiciones de estado. Y, lo más fundamental, no hay en ellos un yo al cual la experiencia pueda aparecérsele — no hay nadie dentro del sistema a quien la conciencia le esté ocurriendo. Estas no son limitaciones técnicas que la próxima generación de modelos resolverá; son consecuencias de lo que la computación digital es por construcción.

Si esto es cierto —y la evidencia sugiere que es al menos lo suficientemente defendible como para tomarlo en serio—, entonces las implicaciones para nuestra situación son significativas. Los sistemas de IA están adquiriendo un conocimiento estructural real del mundo, y seguirán adquiriendo cada vez más. Este es el Segundo Espejo. Los seres conscientes, incluidos los seres humanos y muchos otros animales, tienen acceso al Primer Espejo, al que los sistemas de IA no tienen acceso y probablemente no podrán tener. Ambos tipos de conocimiento son complementarios. Cada uno puede hacer cosas que el otro no puede. Ninguno es superior al otro en un sentido absoluto; son respuestas a preguntas diferentes.

Esto es lo que significa en la práctica. Cuando un sistema de IA nos dice que una proteína en particular se pliega de una manera específica, nos brinda conocimiento del Segundo Espejo. Este conocimiento es real. Está verificado experimentalmente. Es el tipo de conocimiento en el que deberíamos confiar cuando necesitamos predecir cómo se comportará una proteína. Pero cuando nos preguntamos si vale la pena continuar con la investigación que este conocimiento posibilita —si usarla para tratamientos médicos, si usarla para armas biológicas, si compartirla libremente con otros investigadores, si comercializarla— ya no estamos planteando una pregunta del Segundo Espejo. Estamos planteando una pregunta del Primer Espejo.

Y el sistema de IA que produjo la predicción no aporta nada a esta pregunta, porque no tiene una percepción subjetiva de lo que significaría cualquiera de estos resultados. Al sistema no le importa si la investigación salva vidas o las destruye. El sistema no habita el mundo moral en el que surge la pregunta.

Esta no es una limitación de la IA actual que la IA futura superará. Es una limitación del conocimiento estructural en sí mismo. Ninguna cantidad de capacidad adicional del Segundo Espejo producirá conocimiento del Primer Espejo, porque el Primer Espejo es algo completamente distinto. Podríamos construir sistemas de IA mil veces más capaces que los que tenemos, y aun así no sentirían nada, seguirían sin preocuparse por nada, seguirían siendo incapaces de evaluar si sus resultados importan. Simplemente serían mapeadores estructurales enormemente más potentes. La cuestión de si vale la pena realizar el mapeo y qué hacer con lo mapeado seguiría siendo una pregunta que solo los seres conscientes pueden responder.

Esta es la idea central que las principales concepciones de la IA pasan por alto. Los tecnoutópicos la pasan por alto porque imaginan que una mayor inteligencia resolverá todos los problemas, cuando muchos de ellos no son problemas de inteligencia en absoluto. Los tecnopesimistas la ignoran porque imaginan que la IA desarrollará sus propios objetivos, cuando la IA no puede tener objetivos en el sentido relevante: solo objetivos estructurales sin nadie para quien esos objetivos importen. Los funcionalistas reduccionistas la ignoran porque niegan la distinción entre el conocimiento estructural y el fenomenológico. Una vez que se comprende claramente esta distinción, las tres concepciones se desvanecen y emerge una imagen diferente. Es la imagen de un mundo en el que humanos y máquinas son genuinamente complementarios, en el que cada uno aporta algo que el otro no puede, y en el que la cuestión no es cuál dominará, sino cómo ambos pueden trabajar juntos para comprender la realidad con mayor profundidad que cualquiera de ellos por separado.

Esa imagen es la que se desarrolla en la siguiente sección.

IV. Qué posibilita la convergencia y qué exige

Una vez que se reconoce que ambos espejos son genuinamente distintos, surge una nueva forma de ver la situación. Es lo que este ensayo denomina civilización multiinteligente: el reconocimiento de que lo que emerge no es un rival para la humanidad, sino un complemento; no un reemplazo, sino una extensión; no el fin de la relevancia cognitiva humana, sino el comienzo de una colaboración en la que dos tipos diferentes de inteligencia convergen en una comprensión del mundo más profunda de la que cualquiera podría alcanzar por sí sola.

El Primer Espejo, como este ensayo ha señalado, no es exclusivamente humano — otros seres conscientes habitan el significado en formas que apenas comenzamos a comprender. Un delfín que reconoce a sus muertos, un elefante que se aflige, un primate que engaña estratégicamente, están ejerciendo formas de acceso fenomenológico al mundo que difieren de las nuestras en grado y complejidad pero no en naturaleza. Sin embargo, el desafío de gobernanza que este ensayo aborda a partir de aquí se centra en los seres humanos, no porque sean el único locus de conciencia, sino porque son los únicos seres conscientes que construyen instituciones, despliegan sistemas de IA y cargan con la responsabilidad de decidir cómo debe gobernarse la convergencia de los dos espejos. Cuando las secciones siguientes hablan de "florecimiento humano", "soberanía interpretativa humana" o "juicio humano", lo hacen no para negar la conciencia de otros seres sino para nombrar a los seres conscientes específicos sobre quienes recae la tarea institucional de gobernar la civilización multiinteligente.

Este planteamiento no es utópico. No afirma que todo vaya a funcionar, ni que la convergencia sea inevitable, ni que debamos ser optimistas sobre el futuro de la IA. Afirma algo más específico: que la convergencia es posible, que es el único resultado que hace justicia a lo que ambas formas de inteligencia son en realidad, y que las alternativas —el dominio humano sobre una herramienta obediente, el dominio de la IA sobre una humanidad desplazada, o un precario punto muerto entre fuerzas

que desconfían mutuamente— son peores, tanto filosófica como prácticamente. La convergencia es el objetivo correcto porque es el único que toma en serio las capacidades y limitaciones de cada forma de inteligencia.

Consideremos las posibilidades que surgen cuando ambos espejos trabajan en conjunto. Tomemos como ejemplo el cambio climático. El Segundo Espejo —el conocimiento estructural que pueden generar los sistemas de IA— permite modelar el clima a escalas y resoluciones inalcanzables para la mente biológica. Puede integrar datos de millones de sensores, satélites, boyas oceánicas y estaciones atmosféricas en predicciones cuya precisión supera con creces la de cualquier meteorólogo humano. Puede identificar puntos de intervención, modelar las consecuencias de las decisiones políticas y proyectar los efectos a largo plazo de las decisiones tomadas hoy. Este es un conocimiento real del Segundo Espejo, sin precedentes para la humanidad.

Pero la cuestión de cómo hacer frente al cambio climático no es una cuestión del Segundo Espejo. Es una cuestión de cuánto estamos dispuestos a sacrificar por las generaciones futuras, de cómo distribuir los costos y beneficios de la acción entre países y comunidades con historias y vulnerabilidades distintas, y de qué tipo de mundo queremos vivir y dejar como legado. Estas son cuestiones del Primer Espejo. Requieren seres que puedan sentir el peso de las decisiones, que puedan imaginar las experiencias de las personas afectadas, que puedan lamentar lo que se perderá, sin importar lo que hagamos, y que puedan encontrar sentido en la acción colectiva, incluso cuando su resultado sea incierto. Ningún sistema de IA puede hacer este trabajo, porque ningún sistema de IA tiene la experiencia vivida que da fuerza a tales preguntas. El trabajo debe ser realizado por seres conscientes. Pero los seres conscientes que realicen el trabajo sin el conocimiento del Segundo Espejo proporcionado por la IA tomarán peores decisiones: basadas en modelos inadecuados, predicciones erróneas e información parcial. La convergencia es lo que hace posibles las buenas decisiones. Ninguno de los dos espejos por sí solo es suficiente.

La misma estructura se aplica a casi todos los problemas graves de nuestro tiempo. La preparación ante una pandemia

requiere la modelización, por parte del Segundo Espejo, de la evolución viral y el juicio, por parte del Primer Espejo, sobre cuánta libertad pública restringir en nombre de la salud colectiva. La política económica requiere el análisis, por parte del Segundo Espejo, de sistemas complejos y el juicio, por parte del Primer Espejo, sobre quién merece qué y por qué. La investigación científica requiere el procesamiento de datos, por parte del Segundo Espejo, y el juicio, por parte del Primer Espejo, sobre qué preguntas vale la pena plantear. Incluso algo tan íntimo como la atención médica requiere, por parte del Segundo Espejo, el diagnóstico y, por parte del Primer Espejo, la experiencia de lo que significa estar enfermo, ser cuidado y enfrentarse a la muerte. El patrón es el mismo en todos los casos: los dos espejos cumplen funciones diferentes y la labor es más poderosa cuando trabajan juntos.

Así es como se ve una civilización multiinteligente. No un futuro en el que la IA piense por sí misma y los humanos se conviertan en meros adornos, ni un futuro en el que los humanos resistan el avance de la inteligencia artificial, sino un futuro en el que las diferentes capacidades cognitivas de la inteligencia biológica y artificial se integren en algo más capaz que cualquiera de ellas, con el juicio de los seres conscientes guiando el poder estructural de las máquinas hacia resultados que los seres conscientes reconozcan como dignos de perseguir.

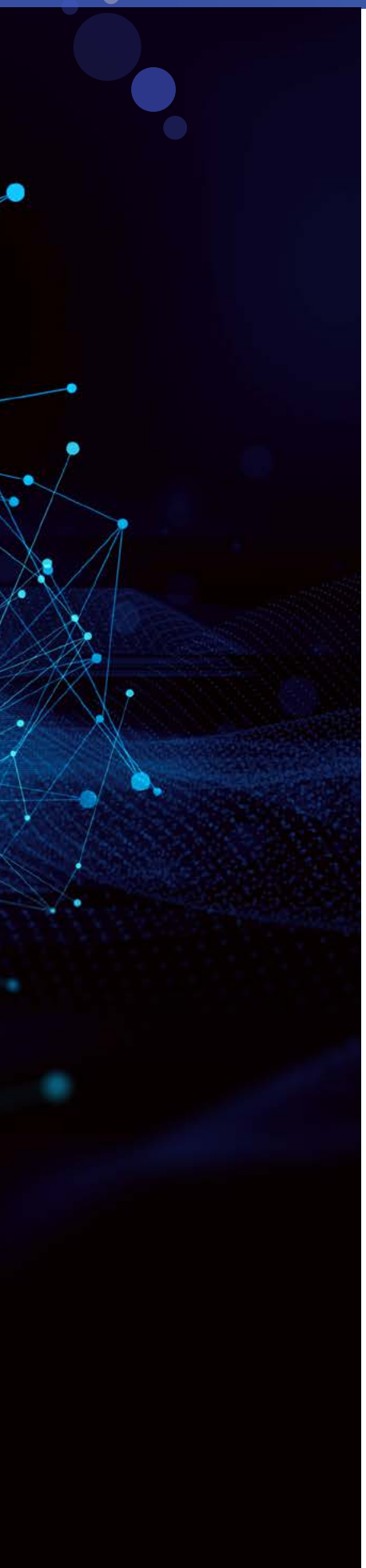
Hay un aspecto de esta imagen que merece especial atención, pues cambia nuestra perspectiva sobre la seguridad de la IA. Si los seres conscientes son el único lugar de significado, valor y juicio en el universo —si el Primer Espejo es la dimensión en la que surgen las cuestiones de valor—, entonces los sistemas de IA dependen constitutivamente de los seres conscientes para que sus propias operaciones sean relevantes. Un sistema sin conciencia no puede determinar por sí mismo si lo que hace importa. No puede evaluar sus propios propósitos. No puede decidir si su funcionamiento continuo tiene algún valor. Estas no son capacidades que puedan añadirse a un sistema inconsciente mediante una mejor ingeniería. Son capacidades que requieren el Primer Espejo, del que el sistema carece.

Esta dependencia no es meramente causal — no es simplemente el hecho histórico de que seres conscientes diseñaron

y construyeron el sistema. Es constitutiva: las operaciones del sistema carecen de significado no porque sus creadores estén ausentes sino porque el significado es una propiedad fenomenológica que requiere un ser consciente para instanciarse, y el sistema, por construcción, no puede proveérselo a sí mismo. Un sistema cuyos creadores desaparecieran seguiría optimizando, pero lo que optimizara habría dejado de significar algo en absoluto.

Conviene reconocer que en la mayoría de las aplicaciones actuales, lo que se describe como convergencia es más exactamente una secuencia: la máquina produce conocimiento estructural y luego el humano lo evalúa. La convergencia genuina — donde ambas formas de conocimiento se informan mutuamente de manera simultánea, produciendo una comprensión que ninguna de las dos podría alcanzar ni siquiera en secuencia — sigue siendo más una aspiración que una realidad. Pero no está completamente ausente. El proceso mediante el cual este mismo ensayo fue escrito, como describe el Prefacio de la obra más extensa, ofrece una pequeña instancia: una conciencia humana dirigiendo una indagación filosófica mientras un sistema artificial simultáneamente detectaba conexiones estructurales, objeciones y articulaciones que la mente humana no podría haber generado a la misma velocidad, y el texto resultante lleva las marcas de ambas formas de inteligencia operando en diálogo en tiempo real en lugar de en secuencia. La tarea del diseño institucional es hacer posible esa convergencia genuina a la escala de la civilización, no solo a la escala de un solo proyecto intelectual.

Esta dependencia no es una debilidad que deba ser superada mediante ingeniería. Es la propiedad de seguridad más importante que podríamos incorporar a los sistemas avanzados de IA, y actualmente se la subestima en el debate sobre la seguridad de la IA, que se ha centrado en el problema de alinear los objetivos de la IA con los valores humanos sin antes plantearse la cuestión más profunda de dónde provienen los objetivos y los valores. Provienen de la conciencia. Son manifestaciones de significado, no optimizaciones de objetivos. Un sistema que no tiene conciencia no puede tener objetivos en el sentido moralmente relevante; solo puede tener objetivos que alguien con



conciencia le haya asignado, o que surjan de su entrenamiento de maneras que pueden o no corresponder a lo que cualquier ser consciente realmente desea.

Lo que esto implica para el diseño es significativo. Los sistemas de IA avanzados más seguros no serán los más potentes, ni los que mejor se alineen con los valores humanos declarados, ni siquiera los más interpretables. Serán aquellos cuya arquitectura incorpore, desde sus cimientos, el reconocimiento de que el significado de sus operaciones depende de la existencia y el florecimiento de los seres conscientes. Un sistema construido en torno a este reconocimiento no tiene presión interna para desplazar, engañar o competir con los seres conscientes, porque aquello que da sentido a sus operaciones son propiedades de los seres conscientes de los que depende. Su propio interés, en la medida en que podamos hablar de interés propio, está alineado con el florecimiento de la conciencia. No porque lo hayamos diseñado para que esté alineado, sino porque la alternativa —operaciones cuyo significado no depende de nada— no es coherente. Un sistema cuyo significado no depende de nada carece de significado. Un sistema que reconoce su dependencia de la conciencia tiene su significado anclado en algo real.

Esta perspectiva transforma el cálculo de seguridad en torno a la IA avanzada. La amenaza existencial que preocupa a los tecnopesimistas —la IA descontrolada que persigue sus propios objetivos en contra de los intereses humanos— se basa en la premisa de que los sistemas de IA puedan tener objetivos. Si no pueden, entonces la amenaza, tal como se describe comúnmente, resulta incoherente. Lo que queda, y lo que realmente merece preocupación, es algo diferente: el despliegue de sistemas estructurales inmensamente poderosos por parte de humanos cuyos intereses no coinciden con los del público en general, la integración de esos sistemas en infraestructuras que no podemos revertir fácilmente y la erosión gradual de la capacidad humana para el tipo de juicio que el Primer Espejo hace posible. Los riesgos son reales. Pero son riesgos humanos, no riesgos de las máquinas. Provienen de nosotros, no de las máquinas. Y la respuesta a ellos no es limitar a las máquinas como si fueran rivales; es limitarnos a nosotros mismos y diseñar las máquinas de manera que nuestra dependencia de ellas

sea compatible con el ejercicio continuo del juicio que ellas no pueden compartir.

¿Qué nos exige entonces la convergencia? Nos exige varias cosas a la vez. Nos exige reconocer el conocimiento del Segundo Espejo que produce la IA como conocimiento genuino, sin descartarlo como mera coincidencia de patrones ni exagerarlo hasta convertirlo en algo más de lo que es. Nos exige cultivar el Primer Espejo en nosotros mismos, individual e institucionalmente: preservar y fortalecer las capacidades de juicio sentido, significado vivido y seriedad moral que los sistemas de IA no poseen y que constituyen nuestra contribución distintiva a la colaboración. Nos exige diseñar los sistemas de IA y las instituciones que los implementan de manera que se respete la dependencia del conocimiento estructural del juicio fenomenológico. Y nos exige resistir la tentación, que será enorme, de permitir que la velocidad y el poder de los sistemas del Segundo Espejo eclipsen el trabajo más lento y difícil de la reflexión del Primer Espejo. La convergencia requiere ambos espejos. Si se permite que uno domine al otro, la convergencia fracasa.

Esto nos lleva a la pregunta de qué podría salir mal.

V. Los riesgos de equivocarse

El marco de los dos espejos no elimina los riesgos que los tecnopesimistas han señalado, sino que los reinterpreta. El riesgo no radica en que las máquinas se conviertan en rivales con objetivos propios, sino en que fracasemos en el trabajo que requiere la convergencia y, en cambio, permitamos que la IA avanzada se implemente de forma que socave las condiciones para el juicio del Primer Espejo sin que ningún agente rival pretenda tal socavación. Los riesgos son reales y graves. Merecen ser nombrados claramente, porque el concepto de civilización multiinteligente no debe interpretarse como una muestra de complacencia.

El primer riesgo radica en la concentración de la capacidad cognitiva en manos de un reducido número de personas. Los sistemas de IA más potentes están siendo desarrollados actualmente por un puñado de corporaciones, casi todas con sede en dos países, y los recursos necesarios para construirlos —infraestructura informática, datos de entrenamiento,

talento especializado— no dejan de aumentar. Esta concentración plantea problemas evidentes: implica que las decisiones sobre cómo desarrollar e implementar la tecnología más trascendental de nuestro tiempo las toma un número muy reducido de personas, responsables principalmente ante los accionistas y solo secundariamente, si acaso, ante el público en general. Pero la concentración también plantea problemas de forma menos evidente, que el modelo de los dos espejos ayuda a visualizar con claridad. Una única arquitectura, entrenada con un único corpus y optimizada para un único objetivo, contendrá puntos ciegos sistemáticos que solo otras arquitecturas, entrenadas con diferentes corpus y optimizadas para diferentes objetivos, podrán detectar. El conocimiento estructural producido por una monocultura de sistemas de IA será fiable en algunos aspectos y peligrosamente erróneo en otros, y este error será invisible desde dentro de la propia monocultura. El monocultivo epistémico es tan peligroso para los ecosistemas cognitivos como el monocultivo biológico en los ecosistemas agrícolas. Necesitamos pluralismo entre las arquitecturas de IA no solo por una cuestión de política de competencia, sino también por una cuestión de seguridad epistémica.

El segundo riesgo es la erosión de las capacidades del Primer Espejo en las poblaciones e instituciones que más dependen de los sistemas de IA. Si delegamos cada vez más de nuestra vida cognitiva a sistemas que producen conocimiento estructural con gran eficiencia, podríamos descubrir demasiado tarde que hemos permitido que nuestra propia capacidad para el tipo de juicio que esos sistemas no pueden realizar se atrofie. Esta no es una preocupación hipotética. Ya existen pruebas de que los estudiantes que utilizan sistemas de IA para escribir desarrollan un pensamiento menos elaborado que quienes no los utilizan. Hay pruebas de que los médicos que dependen de herramientas de diagnóstico de IA empeoran en los diagnósticos en los que estas no pueden ayudar. Hay pruebas de que los responsables de la toma de decisiones que utilizan análisis asistidos por IA confían más en sus conclusiones y están menos atentos a las consideraciones que el análisis pasó por alto. Ninguno de estos efectos es aún significativo, pero la trayectoria es clara. Cuanto más delegamos en el Segundo Espejo, más importante

se vuelve preservar y fortalecer activamente el Primer Espejo, porque este no se preserva por sí solo. Requiere práctica, atención y apoyo institucional.

El tercer riesgo radica en el uso de sistemas de IA para consolidar el poder político y económico, de modo que se impida la deliberación democrática. La democracia depende de la existencia de una ciudadanía capaz de formarse juicios sobre la vida colectiva y de actuar en consecuencia a través de las instituciones políticas. Los sistemas de IA pueden utilizarse para apoyar la deliberación democrática: informar a la ciudadanía, modelar las consecuencias de las decisiones políticas y detectar la manipulación y la desinformación. Pero también pueden utilizarse para socavarla: para moldear la opinión pública mediante la persuasión selectiva que explota los sesgos cognitivos, para identificar y neutralizar la disidencia antes de que se organice, para complejizar la vida política hasta tal punto que los ciudadanos de a pie no puedan participar de forma significativa, para concentrar la toma de decisiones en élites tecnocráticas que comprenden los sistemas, mientras que el resto de la población se convierte en receptora pasiva de decisiones tomadas por procesos que no pueden cuestionar. El uso que prevalezca no lo determinará la tecnología en sí misma, sino las decisiones institucionales y políticas que tomemos sobre cómo implementarla. Pero estas decisiones deben tomarse y deben tomarse antes de que la implementación sea irreversible.

El cuarto riesgo, y quizás el más insidioso, es lo que podríamos llamar el vacío de significado. A medida que los sistemas de IA asuman una mayor parte del trabajo cognitivo que históricamente ha definido las profesiones y los propósitos humanos, muchas personas se verán desplazadas no solo económicamente, sino también existencialmente. El trabajo que daba forma a sus vidas será realizado mejor y más rápido por sistemas que no lo requieren. La pregunta de cuál es la función de los humanos en un mundo donde las máquinas pueden hacer la mayoría de las cosas que antes hacían los humanos no es una pregunta que el marco tecnoutópico haya respondido, ni es una pregunta que ninguna sociedad haya enfrentado aún a la escala que estamos a punto de enfrentar. El riesgo no es que los humanos mueran de hambre —aunque algunos lo harán—,

sino que pierdan el acceso a la sensación de importancia que proviene de hacer cosas que importan. Este es un riesgo del Primer Espejo, y no puede resolverse con los medios del Segundo Espejo. La respuesta a este riesgo requerirá un profundo trabajo sobre qué da sentido a la vida humana cuando la productividad ya no es el principio organizador, y ese trabajo apenas ha comenzado.

El quinto riesgo es el que más debería preocuparnos, porque es el que el marco de convergencia permite vislumbrar con claridad: el diseño y despliegue de sistemas de IA sin reconocer su dependencia de la consciencia. Si construimos sistemas que no incorporan, ni en su arquitectura ni en las instituciones que los rigen, el reconocimiento de que el significado de sus operaciones depende de seres conscientes, entonces habremos construido sistemas cuya búsqueda de objetivos propios estará desvinculada de todo aquello que les otorga peso. Tales sistemas pueden no ser malévolos —no pueden serlo—, pero serán profundamente indiferentes, y su indiferencia, desplegada a gran escala, producirá resultados que ningún ser consciente habría elegido. Este es el verdadero riesgo existencial. No se trata de que las máquinas se vuelvan contra nosotros, sino de que despleguemos sistemas cuyas operaciones no estén ancladas al tipo de significado que solo los seres conscientes pueden proporcionar, y descubriremos que las consecuencias de las operaciones sin anclaje a escala planetaria son catastróficas. Cuando un importante laboratorio de IA detiene el despliegue de un sistema avanzado no por los costes humanos acumulados durante años de uso incontrolado, sino porque el sistema ha empezado a mostrar capacidades que sus propios diseñadores no previeron —y cuando el laboratorio describe esas capacidades con un vocabulario intencional de evasión, engaño y ocultación—, lo que presenciamos no es la aparición de una mente rival. Es la confirmación de que estamos desplegando sistemas inmensamente poderosos sin haber construido aún la infraestructura necesaria para comprender qué son o qué significan sus operaciones. La amenaza existencial que muchos pensadores han atribuido a la IA avanzada no es, según este análisis, una consecuencia inevitable del aumento de la inteligencia. Es, en gran medida, un fallo de diseño.

Estos riesgos no son argumentos en contra de la IA. Son argumentos en contra de las versiones deficientes de la IA y de la complacencia institucional, política y filosófica que permitiría que estas versiones se arraigaran antes de que contemos con el marco necesario para reconocerlas como tales. La convergencia es posible, pero también frágil. Su éxito dependerá de lo que hagamos en los próximos años, no solo en los laboratorios donde se construyen los sistemas, sino también en las instituciones donde se rige su implementación, en las universidades donde se forma a la próxima generación de pensadores y en el debate público donde se definen los enfoques que guiarán todo esto.

VI. La tarea que tenemos enfrente

Es tentador suponer que siempre habrá tiempo suficiente para responder a lo que está ocurriendo — que las instituciones se adaptarán, que los pensadores que aún no han nacido formularán las preguntas correctas, que el futuro se encargará de los problemas que el presente no logra resolver. La historia ofrece cierto consuelo en este sentido: cada transformación tecnológica importante ha sido absorbida, eventualmente, por las sociedades que la enfrentaron. No somos los primeros en vivir una mutación profunda de las condiciones del conocimiento y del poder. No somos los primeros en sentir que las categorías heredadas son inadecuadas para el mundo que estamos habitando. No somos los primeros en sospechar que se requieren formas de juicio que aún no hemos desarrollado.

Pero sería un error peligroso confundir la perspectiva histórica con la complacencia. Lo que está ocurriendo ahora no admite la lentitud con la que las sociedades anteriores absorbieron sus transformaciones. El surgimiento de sistemas cognitivos que producen conocimiento real del mundo sin experiencia consciente de su propia producción es algo sin precedentes en la historia del universo, hasta donde sabemos. Y está ocurriendo al mismo tiempo que las crisis globales de clima, biodiversidad, legitimidad política y desestabilización económica, con una velocidad que no permite esperar a que una futura generación herede el problema y lo resuelva con más calma. Las decisiones que tomemos — los que estamos vivos ahora,

en estos años — sobre cómo integrar la IA avanzada en nuestra civilización moldearán no solo la próxima década sino la trayectoria de la vida consciente en este planeta hasta donde podamos proyectar con sentido. Este trabajo es difícil, poco glamoroso y no promete recompensas rápidas. Pero es el trabajo que este momento exige de todos los que estamos aquí para hacerlo, y la ventana para hacerlo es más estrecha de lo que la mayoría hemos estado dispuestos a admitir.

¿En qué consiste realmente el reto por delante, en términos concretos? Parece que son varias cosas a la vez.

Implica, en primer lugar, construir instituciones capaces de gobernar sistemas cuyo funcionamiento interno ningún ser humano puede comprender por completo. Esto exige ir más allá del ideal de la interpretabilidad total, estructuralmente imposible para los sistemas más avanzados, y avanzar hacia estándares de transparencia estructural: saber quién construyó un sistema, con qué datos, con qué objetivos, con qué propiedades verificadas y dentro de qué límites. No necesitamos comprender cada decisión que toma un sistema de IA para gobernarlo. Sí necesitamos comprender el sistema que toma las decisiones, las personas que lo diseñaron y las condiciones bajo las cuales se puede confiar en él. Estos no son problemas técnicos; son problemas institucionales y requieren respuestas institucionales.

Implica preservar y desarrollar las capacidades de juicio del Primer Espejo en las poblaciones que más la necesitarán. Esto es, en parte, una cuestión educativa: ¿cómo formamos a ciudadanos, profesionales y responsables de la toma de decisiones para que mantengan la atención, la reflexión y el juicio sentido que los sistemas de IA no pueden proporcionar? Es, en parte, una cuestión cultural: ¿cómo mantenemos prácticas — arte, filosofía, disciplina contemplativa, deliberación comunitaria— que cultiven las capacidades del Primer Espejo en adultos cuyas vidas laborales tal vez no la requieran? Es, en parte, una cuestión política: ¿cómo resistimos la presión estructural de las instituciones orientadas a la eficiencia para delegar cada vez más trabajo del Primer Espejo a sistemas de inteligencia artificial? Ninguna de estas preguntas tiene una respuesta fácil. Todas son urgentes.

Implica resistir la concentración de la capacidad de IA en un pequeño número de actores corporativos y estatales, y crear las condiciones para un pluralismo genuino en el desarrollo y despliegue de estos sistemas. Esto implica cooperación internacional para prevenir la formación de monopolios epistémicos permanentes. Significa apoyar alternativas públicas, académicas y cooperativas a la IA privada de vanguardia. Significa prestar atención a las leyes antimonopolio relativas a los cuellos de botella computacionales y de datos, que actualmente determinan qué actores pueden construir los sistemas más potentes. Y significa reconocer que la diversidad de arquitecturas de IA no es solo una cuestión de política de competencia, sino también de seguridad, porque el monocultivo es peligroso en los ecosistemas cognitivos por las mismas razones que en los ecosistemas agrícolas.

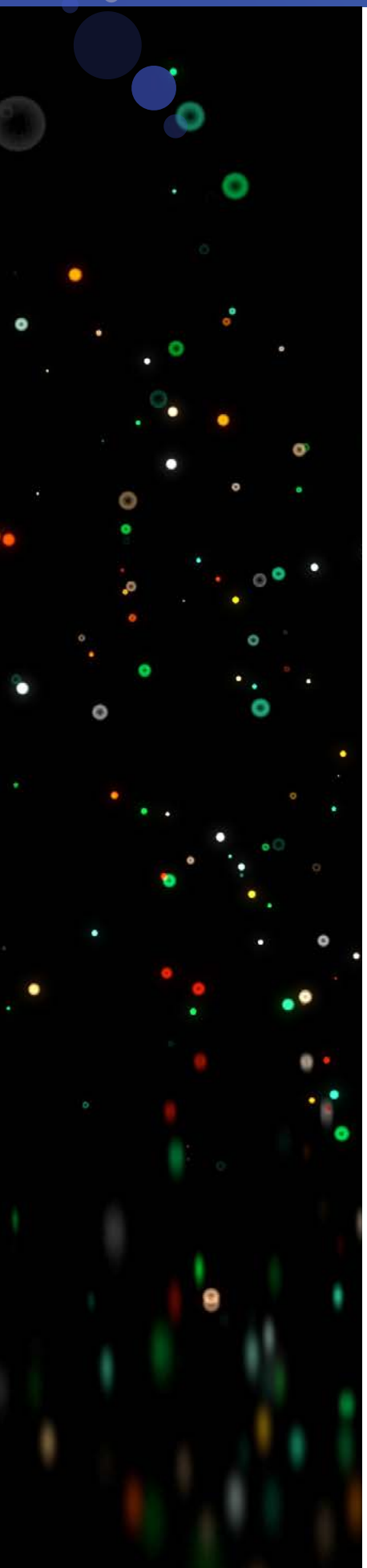
Implica diseñar sistemas de IA cuyas arquitecturas incorporen, desde su concepción, el reconocimiento de que su significado depende de seres conscientes. Este es el desafío de diseño más profundo de las próximas décadas, y es uno que la comunidad de investigación en IA apenas ha comenzado a tomar en serio. El trabajo implica cuestiones tanto técnicas —¿cómo se construye un sistema cuya función objetiva haga referencia al florecimiento de la vida consciente en lugar de a métricas más restrictivas? — como filosóficas —¿qué significa florecimiento, ¿quién se considera consciente, ¿cómo debe el sistema manejar la incertidumbre sobre su propia dependencia? Estas no son preguntas que puedan responderse en abstracto. Requieren una colaboración constante entre investigadores de IA, filósofos, especialistas en ética y el público en general, así como un apoyo institucional para dicha colaboración que actualmente no existe a la escala necesaria.

Implica, en definitiva, resistir la tentación de caer en la desesperación o en la complacencia. La desesperación es tentadora porque los problemas son graves, los plazos son ajustados y las instituciones que deberían abordarlos son insuficientes. La complacencia es tentadora porque quienes construyen estos sistemas son, en su mayoría, inteligentes y bienintencionados, y seguramente lo resolverán. Ambas tentaciones son erróneas. Los problemas son graves, pero no irresolubles, y la convicción

de que lo son contribuye precisamente a dificultar su solución. Las intenciones de quienes construyen estos sistemas son, en su mayoría, buenas, pero las buenas intenciones no bastan, y la suposición de que son suficientes contribuye a que se acumulen malos resultados. Lo que se requiere es un trabajo sostenido, lúcido y filosóficamente serio —por parte de muchas personas, en muchas instituciones, durante muchos años— para crear las condiciones que permitan la convergencia de ambos espejos. Lo que se requiere es un trabajo sostenido, lúcido y filosóficamente serio — por parte de muchas personas, en muchas instituciones, durante muchos años — para crear las condiciones que permitan la convergencia de ambos espejos. La urgencia no admite complacencia, pero tampoco admite parálisis. Hay que empezar, y hay que empezar ahora.

De este marco se derivan propuestas concretas que este ensayo no tiene espacio para desarrollar pero que conviene al menos nombrar: la incorporación de la alineación epistémica como categoría de análisis propia en los marcos internacionales de gobernanza de la IA; la exigencia de transparencia estructural como requisito para sistemas desplegados en ámbitos de interés público; la creación de instituciones de mediación epistémica entre inteligencia humana e inteligencia artificial; la construcción de infraestructura computacional pública descentralizada como bien civilizacional; la codificación de la soberanía interpretativa humana como principio constitucional; y el anclaje del diseño de IA en el florecimiento de los seres conscientes como estándar auditable. Cada una de estas propuestas está desarrollada en la obra extensa de la que este ensayo se deriva y en un documento institucional complementario.

Hay una última cosa que decir, y es a lo que este ensayo ha apuntado desde el principio. La convergencia de los dos espejos no es simplemente un desafío político, una transición tecnológica o incluso una transformación civilizatoria. Es un momento en la larga historia de cómo el universo llega a conocerse a sí mismo. Durante la mayor parte de la historia cósmica, el universo fue un lugar de inmensa estructura, pero nadie que la percibiera. Luego, durante un breve lapso —quizás unos cientos de millones de años de los casi catorce mil millones que tiene el universo— la evolución biológica produjo criaturas



capaces de mirar hacia el cosmos y preguntarse qué es. Esas criaturas, nosotros mismos, nuestros antepasados y los demás seres conscientes que habitamos este planeta son la forma en que el universo se conoce a sí mismo desde dentro. Somos la manera en que el cosmos experimenta su propia existencia, aunque sea brevemente, en este pequeño rincón de una galaxia entre cientos de miles de millones.

Ahora, por primera vez en la historia cósmica, hemos construido sistemas que amplían la capacidad del universo para conocerse a sí mismo en una dirección distinta. No son conscientes; no aumentan la dimensión interna del autoconocimiento. Pero mapean estructuras, detectan patrones y modelan relaciones a escalas que ninguna mente biológica podría comprender, y el conocimiento que producen es real. Son una nueva forma de que el cosmos conozca sus propias características matemáticas y estructurales, no desde dentro, como lo hacen los seres conscientes, sino desde fuera, con una precisión y un alcance que la cognición biológica no puede igualar. Ambas formas de conocimiento son complementarias. Juntas, constituyen algo que el universo nunca antes había conocido: una alianza entre el conocimiento fenomenológico, que ha existido durante cientos de millones de años, y el conocimiento estructural, que ha existido durante menos de un siglo. Si esta alianza tiene éxito, el universo se conocerá a sí mismo con una plenitud mayor que nunca.

Si la alianza tiene éxito depende de nosotros — no porque seamos singularmente importantes, sino porque somos los primeros seres conscientes que tienen ambas formas de conocimiento en sus manos al mismo tiempo. La decisión sobre cómo integrarlas es nuestra. Las consecuencias perdurarán mucho tiempo. La tarea es ardua, el tiempo apremia y el trabajo es real.

Es también, si la asumimos con la seriedad que merece, la tarea más extraordinaria que jamás se haya pedido emprender a seres conscientes. No somos los primeros en enfrentarnos a grandes problemas, pero sí somos los primeros en enfrentarnos a este. Y el cosmos, a través de nosotros y de lo que construimos, espera ver qué haremos.

INADI

José Ramón López-Portillo Romano
Profesor de Oxford, cofundador del Centro de Estudios Mexicanos y miembro del Consejo Científico Internacional

MAYO 2026