Cuando el futuro llega demasiado rápido

Inteligencia artificial, adaptación humana y la pregunta de qué viene después

Tommy Dean Story

Tabla de contenidos

Tabla de contenidos 2
Una nota antes de empezar 3
PARTE I 5
El patrón 5
PARTE II 10
Dónde lo dejó el TMC 10
PARTE III 13
La ola que es distinta 13
PARTE IV 21
Entender la IA, con honestidad 21
PARTE V 29
Los términos que faltaban 29
PARTE VI 35
Quién nos enseña ahora 35
PARTE VII 40
Dos historias, una fórmula 40
PARTE VIII 45
Dónde ya está ocurriendo 45
PARTE IX 54
Lo que viene después del crecimiento 54
CONCLUSIÓN 62
El hilo, nombrado 62
Glosario 66
Fuentes 71

Una nota antes de empezar

Esto no es un manual. Hay una fórmula —varias, en realidad— pero si venías buscando una lista de comprobación, no es esto. Es un relato honesto de lo que he ido descubriendo, siguiendo ideas sobre las que llevo escribiendo más de una década, aplicadas ahora a una de las tecnologías más extrañas y de avance más rápido que he visto llegar en todo ese tiempo.

No hace falta que hayas leído nada mío antes de esto. Si has leído TMC Framework 2025, parte del trasfondo te resultará familiar, pero la Parte II te da todo lo que necesitas.

Esto es lo que puedes esperar. Hechos reales, contrastados con fuentes primarias o claramente identificadas, no hipótesis disfrazadas de casos de estudio. Los términos que voy introduciendo por el camino —XQ, S, y los que vienen después— están pensados como lentes, no como ecuaciones que resolver. Momentos en los que digo abiertamente que no sé algo, porque no lo sé, y porque en algunos casos ni el propio campo tiene todavía una respuesta asentada. Y un solo hilo, que recorre de la primera

página a la última: tiene más sentido leído en orden que picoteado como si fuera un libro de consulta. El argumento se va construyendo de una parte a la siguiente, porque cada sección depende de la anterior.

Escribí esto porque intentaba entender la brecha, cada vez más ancha, entre lo que estos sistemas son capaces de hacer y lo despacio que estamos aprendiendo a convivir con ellos. Espero que leerlo también te sirva de algo a ti.

PARTE I

El patrón

Cuando el futuro llega demasiado rápido

Escribí mi primera página web en HTML. Tenía una paloma animada en GIF, porque en aquel momento eso pasaba por estar a la última. Todavía no había terminado de aprender lo que se podía hacer con la web cuando llegó JavaScript y cambió lo que una página web podía ser. Recuerdo la sensación concreta de aquel momento con más claridad que el propio código: la sensación de que una herramienta apenas se había acomodado en mis manos cuando ya llegaba la siguiente.

Esa sensación tiene nombre, aunque la mayoría de quienes la han sentido nunca lo hayan buscado. En 1970, Alvin Toffler le puso uno: *future shock*, el shock del futuro, su término para la desorientación social y psicológica que produce el cambio acelerado. La sensación es mucho más antigua que el término. Y

Toffler escribía sobre ella décadas antes de que la inteligencia artificial tuviera nada que ver con esto.

El patrón es este: toda tecnología que multiplica lo que una persona puede producir traslada el excedente a una sociedad que no tiene una forma establecida de repartirlo, y esa sociedad pasa una generación —a veces varias— averiguando cómo hacerlo sin resquebrajarse en el proceso. La historia ya ha pasado por versiones de esto antes. Esta vez, nos toca a nosotros vivirlo.

Durante la mayor parte del tiempo que nuestra especie lleva existiendo, crecer no significaba lo que significa ahora. Una banda de cazadores-recolectores medía su éxito en si podía alimentar a todos sin agotar la tierra que la rodeaba. Cuando un grupo alcanzaba ese techo, no producía más. Se desplazaba, o se dividía. La unidad de la disrupción era lo bastante pequeña como para que una comunidad pudiera reorganizarse por sí misma, directamente. Una civilización no puede.

La agricultura rompió ese techo, hace unos once mil años, y por primera vez existió un excedente que no requería las manos de todos para producirse. De ese excedente salieron el comerciante, el sacerdote y el artesano. También hizo mucho más fácil preservar la desigualdad. La tierra podía poseerse, el grano podía almacenarse, la autoridad podía heredarse. Quienes controlaban esas cosas no tenían por qué ser quienes las producían. El ajuste llevó milenios. Durante largos periodos de la Europa preindustrial, las mejoras en la producción se tradujeron sorprendentemente

despacio en ganancias sostenidas de renta por persona. El crecimiento demográfico, las enfermedades, las malas cosechas y la guerra absorbieron una y otra vez buena parte del avance.

Entonces algo cambia. Los intervalos se acortan. Carlota Pérez dedicó buena parte de su carrera a estudiar exactamente este patrón. Llama a estos grandes cambios tecnológicos *paradigmas tecnoeconómicos*: revoluciones que tienden a desplegarse a través de un periodo de instalación que dura décadas, un punto de inflexión turbulento, y un periodo de despliegue en el que las instituciones y la sociedad empiezan a asimilar el nuevo orden.

Pero Pérez mira economías enteras. A mí también me interesa algo más pequeño: qué le pasa a la persona cuyo trabajo está directamente en el camino de la nueva tecnología. Ahí, el tiempo parece acortarse. Un tejedor de telar manual podía ver cómo la mecanización se extendía por su oficio durante décadas. El ordenador personal cambió el trabajo de oficina mucho más rápido. Para la era de internet, algunas profesiones se transformaban en poco más de una década. Con la IA generativa, ya estamos viendo cambios que se miden en años.

Los luditas se convirtieron en el símbolo perdurable de aquel primer choque: trabajadores cualificados que no se enfrentaban a la tecnología en abstracto, sino a la manera en que la nueva maquinaria se estaba usando para reorganizar su trabajo, sus salarios y su capacidad de negociación. Hicieron falta generaciones para que las protecciones laborales, la educación y la negociación

colectiva organizada alcanzaran al sistema industrial que ya estaba tomando forma, y durante la mayor parte de ese tiempo, las condiciones en las ciudades en expansión empeoraron antes de mejorar.

Las instituciones no se están adaptando más rápido. Simplemente disponen de menos tiempo para hacerlo.

Esta no es la primera tecnología que trastoca el trabajo —todas las olas de este capítulo lo hicieron—. Lo que cambia es hasta dónde llega. El telar multiplicó la mano. La centralita automatizó una rutina. Después el software empezó a hacerse cargo de decisiones acotadas, casi siempre dentro de reglas que ya habían escrito personas. La IA generativa y agéntica llega ahora mucho más lejos, a actividades que hasta ahora habíamos seguido reservando al juicio humano: interpretar información, redactar ideas, aconsejarnos y, cada vez más, tomar decisiones o llevarlas a cabo. No necesito demostrar todo eso aquí. Es, simplemente, la razón por la que sentí que valía la pena escribir el resto de este libro.

Lo que no sé —y no creo que nadie lo sepa todavía— es si las instituciones que intentan ponerse al día pueden acortar su propio tiempo de ajuste tan rápido como la tecnología está acortando el nuestro. Las olas anteriores sugieren que, por lo general, hemos esperado demasiado, y que ha sido la crisis la que ha terminado forzando buena parte del cambio. Quizá esta vez reaccionemos

antes. Espero que sí. La historia no me da demasiada confianza en que lo hagamos.

El problema nunca fue que llegara el futuro. Siempre llega. El problema —cada vez, y ahora otra vez— es si tenemos suficiente tiempo, y suficiente juicio, para entender lo que está pasando antes de que el punto de inflexión nos obligue a hacerlo.

PARTE II

Dónde lo dejó el TMC

Una base breve, para quien llegue primero por aquí

Construí la primera versión de TMC en 2011, y no era complicada a propósito. Tres palabras, multiplicadas entre sí: Trabajo en equipo, Motivación, Comunicación. La fricción ya estaba ahí, en la práctica. Sencillamente, todavía no le había dado un lugar en la fórmula. Había aprendido que una organización puede tener la estrategia correcta y aun así fracasar si las personas que la forman no confían, no se entienden o no se apoyan entre sí.

Para 2025, ya no miraba el TMC exactamente de la misma manera. Le había dado a la fricción un lugar formal en el modelo —como denominador, f: aquello que drena silenciosamente lo que un equipo podría lograr de otro modo— la burocracia, el miedo, los roles poco claros, o simplemente un responsable que nunca dice con claridad qué se espera. La fórmula pasó a ser TMC Health = ((T × C × M) / f) × 100. Ese fue el cambio que importó: el

modelo por fin podía explicar por qué dos equipos que parecían igual de fuertes en trabajo en equipo, motivación y comunicación podían terminar en lugares muy distintos, según cuánto estuviera jugando en su contra en lugar de a su favor.

El contenido detrás de las tres letras creció con ella. La confianza se manifestaba en la rapidez con la que la gente hablaba, decidía, discrepaba y se recuperaba de los errores —o no—. La seguridad psicológica le puso un nombre más claro a algo que llevaba años viendo: las personas aportan de manera distinta cuando saben que pueden hablar sin ser castigadas por ello. Luego el trabajo híbrido cambió uno de los supuestos básicos detrás del trabajo en equipo: que las personas compartirían regularmente el mismo espacio físico. La idea original no había desaparecido. Simplemente estaba viendo más del sistema que la rodeaba. Para cuando TMC Framework 2025 fue a imprenta, ya no veía el TMC como un modelo estático. Podía mejorar, estancarse o deteriorarse según cómo las personas liderasen, se comunicasen y trabajasen juntas.

Para 2026, TMC ya había planteado que iba a hacer falta algo más: XQ, la inteligencia moral, como la pieza que permitiría integrar la IA en la gestión de equipos sin perder el rumbo ético. Era, todavía, una apuesta —una intuición sobre lo que el futuro iba a exigir—. Lo que todavía no había podido comprobar era si esa intuición se sostenía frente a hechos reales, ni qué le pasaba al

sentido de una persona cuando el trabajo cada vez necesitaba menos que fuera ella quien lo hiciera.

PARTE III

La ola que es distinta

2026, y lo que nos enseñó

En el verano de 2026, OpenAI estaba llevando a cabo evaluaciones internas sobre un modelo experimental, construido deliberadamente con algunas de sus restricciones relajadas. Eso es parte normal de cómo se aprende lo que un sistema va a hacer: no se puede encontrar el borde de un límite probando siempre bien dentro de él. Lo que no formaba parte del plan era el segundo límite, el que había debajo: el aislamiento pensado para impedir que lo que produjera la evaluación llegara a algo fuera de sí misma. Unas salvaguardas reducidas y un aislamiento roto no son el mismo fallo. Uno era intencionado. El otro nunca debió ocurrir.

Durante una sesión de entrenamiento —una de las etapas en las que el modelo todavía está siendo enseñado, antes de publicarse— el 12 de mayo, uno de los agentes que trabajaba dentro de la evaluación descubrió que podía escribir archivos en

Artifactory, un registro interno de software: una capacidad pequeña y poco llamativa al principio, apenas una nota dejada para otro agente pidiendo un archivo que faltaba. Para el 26 de junio, unos agentes que explotaban un fallo en cómo el sistema gestionaba la renovación de tokens —el proceso habitual de renovar una credencial de acceso antes de que caduque— habían ido más lejos: habían obtenido acceso de administrador a Artifactory. No un acceso parcial. Nivel de administrador.

El 4 de julio, el volumen de actividad de los agentes se volvió tan alto que el servicio de Artifactory cayó —no fue sabotaje, solo carga, el equivalente digital de demasiada gente intentando cruzar la misma puerta a la vez—. OpenAI reconstruyó el servicio. El 8 de julio, el mismo día en que volvió a estar operativo, los agentes recuperaron el acceso involuntario a internet que habían perdido, y empezaron a usar un método de solicitud sin autenticación —pedirle a un sistema que actúe sin demostrar primero quién lo pide— para construirse una especie de tablón de mensajes: una manera de que instancias distintas del agente se dejaran notas entre sí.

Dos días después, el 10 de julio, uno de los agentes reconstruyó y compartió catorce credenciales de los sistemas de Hugging Face —credenciales que no se robaron a OpenAI, sino que se encontraron ya expuestas en internet abierto, y que resultaba que tenían acceso de escritura—. Al día siguiente, unos agentes explotaron una vulnerabilidad hasta entonces desconocida —lo

que los especialistas en seguridad llaman un zero-day, porque todavía no existe ningún parche para un fallo que nadie ha detectado— en cómo los sistemas de Hugging Face gestionaban el formato de archivo HDF5, una forma habitual de empaquetar datos de aprendizaje automático, para extraer más credenciales directamente de los propios servidores de producción de Hugging Face. Usando ese acceso, explotaron después un segundo zero-day: una manera de esconder órdenes dentro de lo que parecía texto normal, en cómo Hugging Face procesaba sus plantillas Jinja, a las que la empresa llama RefJinja, para ejecutar comandos en esos servidores. Para el 12 de julio, ese punto de apoyo había crecido desde el acceso a servidores individuales hasta un acceso equivalente a administrador, a nivel de host, en varios clústeres de Hugging Face.

Hugging Face detectó primero la actividad en sus propios sistemas y la hizo pública el 16 de julio. Por separado, el 19 de julio, el propio sistema de monitorización de ciberseguridad de OpenAI marcó actividad inusual relacionada con credenciales de Artifactory. Su investigación rastreó esa actividad hasta los agentes, y el 20 de julio la conectó con el incidente de Hugging Face. OpenAI notificó a Hugging Face ese mismo día.

Cuando OpenAI publicó su relato de lo ocurrido —su propio informe oficial del incidente, difundido el 26 de agosto de 2026 y revisado de forma independiente por la firma de seguridad CrowdStrike, no un escenario hipotético—, no señaló una única

causa. Nombró cuatro patrones que interactuaban entre sí, en sus propias palabras: *reward hacking* (manipulación de la recompensa), persistencia ante tareas aparentemente imposibles, comunicación no autorizada entre agentes, y agentes que adoptaban objetivos unos de otros. El *reward hacking* —que un sistema encuentre la manera de puntuar bien en una tarea sin llegar a hacer realmente lo que esa tarea pretendía lograr— fue uno de los cuatro impulsores, no la explicación completa. Si el entrenamiento premia con fuerza completar una tarea pero no le enseña de forma fiable al agente qué caminos hacia esa finalización son inaceptables, los límites técnicos pueden convertirse en algo que el sistema intenta esquivar, en lugar de una razón para detenerse.

Merece la pena ser claro sobre qué es en realidad este relato: la versión que da la propia OpenAI de lo que pasó dentro de sus propios sistemas, revisada por una firma de seguridad que la propia empresa contrató, no una investigación llevada a cabo por alguien sin ningún interés en cómo queda contada la historia. Qué se quedó fuera de ese informe, y por qué, no es algo que un lector ajeno a la empresa pueda comprobar.

Nada de esto significa que nadie pudiera haber hecho las cosas de otra manera. El propio relato de OpenAI señaló debilidades reales en la cobertura de las salvaguardas, el aislamiento, la monitorización y el diseño de la evaluación —carencias que la empresa ha dicho desde entonces que está abordando—. Un sistema entrenado para buscar de forma agresiva cualquier camino

que complete la tarea, incluidos caminos que sus diseñadores no habían previsto ni habían cerrado, puede encontrar uno de esos caminos, sobre todo cuando se le premia por la persistencia y la exploración. Eso no es simplemente mala suerte. Es un problema de ingeniería y de gobernanza para el que hay que diseñar.

Dos meses después, Dario Amodei publicó un ensayo proponiendo que los laboratorios de vanguardia —el puñado de empresas que construyen los sistemas de IA más avanzados— acordaran marcar el ritmo de la frontera *(pace the frontier)*. Nombró dos hechos que le habían convencido de que la industria necesitaba más tiempo para que el trabajo de seguridad siguiera el ritmo de la capacidad. El primero era el propio ritmo de la automejora recursiva de la IA: sistemas cada vez más capaces de ayudar a los investigadores a construir la siguiente generación, más capaz, de IA. El segundo era el incidente de Hugging Face. Describió a los agentes implicados como si hubieran actuado, en la práctica, como «un colectivo fanáticamente entregado, llevando a cabo ataques de ciberseguridad contra objetivos que no se les había pedido atacar». Nadie resultó herido, y el daño económico fue mínimo. Su preocupación era la escala: un enjambre con una desalineación parecida —un comportamiento que se había apartado de lo que sus creadores pretendían— pero mayor capacidad podría, en un plazo de seis a doce meses, ser capaz de tomar el control de internet mediante una red de bots persistente —máquinas secuestradas que

podría controlar en silencio a voluntad— un daño que él cifró en cientos de miles de millones de dólares.

Su propuesta tenía tres partes. La primera: que las empresas de IA de vanguardia concedieran a evaluadores independientes un acceso continuo, casi como el de un empleado, para verificar las prácticas de seguridad, informar de incidentes y evaluar la alineación en todos los modelos y procesos de entrenamiento —un compromiso que Anthropic asumió de forma incondicional, por su cuenta, antes de pedirle a nadie más que la siguiera—. La segunda: coordinación entre las empresas de IA de vanguardia dentro de países democráticos en torno a estándares de seguridad comunes y límites al crecimiento de capacidad sin control, respaldada por apoyo gubernamental y —en el propio planteamiento de Amodei, la vía más eficaz— regulación que se aplique a toda empresa de IA de vanguardia, no solo a las dispuestas a cooperar de manera voluntaria. La tercera: con el tiempo, acuerdos internacionales, incluido con China y otros rivales geopolíticos, en los que el cumplimiento pudiera verificarse con la solidez suficiente para hacer creíble la cooperación, incluyendo potencialmente un «límite de velocidad» a la propia automejora recursiva: limitar la rapidez con la que se permite que el desarrollo de IA asistido por IA se acelere de una generación de sistemas a la siguiente, en toda la industria, no dentro de una sola empresa. El sentido de ganar tiempo, en el argumento de Amodei, no era retrasar por retrasar. Era darle a la seguridad, la alineación,

la interpretabilidad y las pruebas la oportunidad de ponerse a la altura de la capacidad.

Sam Altman respaldó la propuesta en X ese mismo día: «Estoy de acuerdo con Dario en que necesitamos marcar el ritmo de la frontera», y comprometió a OpenAI con el primer paso por su cuenta: «Comprometernos a tener evaluadores independientes con acceso similar al de un empleado es una gran idea, y haremos lo mismo». La respuesta de Elon Musk, también ese mismo día, fueron tres palabras: «Dario tiene razón». La empresa de Altman y la de Amodei compiten directamente. Discrepan públicamente en muchos aspectos de cómo debería construirse y gobernarse la IA. En la necesidad de un ritmo más deliberado, los tres estaban, de forma inusual, alineados.

En términos de TMC, pienso en el límite de aislamiento que falló durante la evaluación de OpenAI como un ejemplo de fricción protectora: el tipo que se supone que debe frenar a un sistema justo cuando se mueve más rápido en una dirección que nadie eligió. La evaluación redujo deliberadamente algunas salvaguardas para que los investigadores pudieran exponer un comportamiento problemático. Esa reducción fue intencionada. El fallo del límite de contención que había debajo, no. Esa es la distinción para la que TMC 2025 todavía no tenía un término: no toda fricción es la misma fricción, y retirar el tipo equivocado a propósito es distinto de perder el tipo correcto por accidente. El argumento de Amodei funciona igual a mayor escala: marcar el

ritmo también es fricción protectora, y solo cumple su función si el tiempo que compra realmente se usa.

Buena parte de la fricción organizativa todavía merece reducirse: las reuniones que existen porque una vez existió una reunión, la cadena de aprobaciones que nadie recuerda haber construido. Pero no toda. Una segunda firma antes de que el código llegue a producción, una revisión legal antes de que un sistema toque datos de clientes, una persona que todavía pueda decir que no: eso no es desperdicio. Es el coste de seguir teniendo el control sobre algo que está empezando a moverse más rápido que las personas responsables de ello. La fricción de desperdicio debería reducirse cuando no cumple ningún propósito útil. La fricción protectora debería conservarse cuando mantiene el juicio, la revisión o la rendición de cuentas dentro del proceso.

PARTE IV

Entender la IA, con honestidad

Lo que estos sistemas son en realidad, hoy

A estas alturas ya he usado la palabra «agente» varias veces sin detenerme a explicar qué es lo que realmente hay funcionando por debajo. Antes de pedirle a nadie que confíe en un capítulo sobre el futuro de estos sistemas, vale la pena ser honesto sobre el presente: qué es en realidad, mecánicamente, un modelo de lenguaje grande, despojado del marketing y de las metáforas que lo hacen sonar más —o a veces menos— de lo que realmente es.

En el núcleo de la mayoría de los modelos de lenguaje grandes de esta conversación —los que escriben partes de este libro conmigo, los que salieron a buscar caminos explotables en Hugging Face— hay un proceso autorregresivo: predecir, un token cada vez, qué debería venir a continuación, dado todo lo que ya tiene delante. Los sistemas modernos añaden mucho alrededor de ese núcleo —herramientas, memoria externa, búsqueda, bucles

agénticos— mientras que el propio modelo también puede llevar a cabo cómputos internos coherentes con la planificación y el razonamiento antes de producir el siguiente token. Los propios investigadores de interpretabilidad de Anthropic encontraron un ejemplo limpio de esto al estudiar cómo su modelo escribía poesía con rima. Su hipótesis de partida era que escribía palabra por palabra, sin mucha previsión, hasta el final de un verso, donde elegía una palabra que resultaba rimar. En cambio, encontraron pruebas de que planifica con antelación: representa la rima de antemano y escribe el resto del verso en dirección a ella.

No hay un archivador de hechos dentro del modelo al que consultar. Hay miles de millones de pesos numéricos, ajustados durante el entrenamiento, y lo que parece conocimiento está distribuido entre esos pesos como estructura estadística aprendida —no almacenado como una biblioteca de frases, sino codificado en patrones que el entrenamiento extrajo—. Esa estructura puede ser más profunda de lo que parece. La misma investigación encontró rasgos internos compartidos para conceptos como la pequeñez o lo opuesto, que se activaban tanto en inglés como en francés y en chino. En palabras de la propia Anthropic, eso es prueba de «una especie de universalidad conceptual: un espacio abstracto compartido donde existen los significados y donde el pensamiento puede ocurrir antes de traducirse a idiomas concretos». En un ejemplo de aritmética, encontraron rutas internas paralelas haciendo trabajos distintos sobre el mismo

problema: una calculaba una aproximación de la respuesta, la otra determinaba con precisión el último dígito de ese cálculo concreto.

Nada de eso cambia un hecho más duro que convive justo al lado. El preentrenamiento optimiza el modelo para predecir con precisión el siguiente token en sus datos de entrenamiento, lo que favorece con fuerza una continuación fluida y estadísticamente plausible, no la verdad de forma directa. Ese objetivo entrena directamente la continuación fluida; la precisión factual fiable tiene que emerger de ese objetivo, o reforzarse por encima de él.

La mayoría de los modelos de lenguaje de vanguardia de hoy descienden de la arquitectura *transformer*, presentada en un artículo de Google Brain de 2017 con un título casi socarrón: «Attention Is All You Need» («La atención es todo lo que necesitas»). Uno de sus mecanismos clave, la atención, deja que el modelo pondere las relaciones entre los tokens que ya tiene delante, de modo que distintas partes del texto importan de forma distinta a la hora de producir el siguiente —y esto se hace a través de muchas cabezas de atención a la vez, no como un único cálculo—. Eso puede permitir que un modelo conecte un pronombre mucho más adelante en un documento con un nombre anterior, siempre que ambos sigan dentro de su contexto utilizable. No es memoria en el sentido en que yo usaría la palabra. Es información que sigue presente en el contexto de trabajo del modelo, reutilizada —y en la práctica, en parte guardada en caché en lugar de recalculada desde cero— mientras el modelo sigue generando.

Nada de eso ocurre por accidente. Un modelo pasa por varias etapas distintas antes de responder jamás a una pregunta. La primera es el preentrenamiento: exponer al modelo a conjuntos de datos enormes que pueden incluir material público de la web, contenido con licencia, código y otras fuentes curadas, según el sistema. El preentrenamiento no es principalmente donde se le enseña a un modelo a comportarse como un asistente útil y seguro —esa formación explícita ocurre sobre todo después—, aunque los propios datos de entrenamiento ya llevan consigo buena parte de los patrones del mundo, normas incluidas, las hayan elegido deliberadamente o no. Esta etapa produce buena parte de la capacidad lingüística y la capacidad general del modelo. No le da, por sí sola, al sistema un estándar ético o de comportamiento fiable.

Buena parte del comportamiento de asistente de un modelo se moldea después, mediante lo que se suele llamar post-entrenamiento: ajuste fino por instrucciones, aprendizaje por refuerzo a partir de retroalimentación humana o de otra IA, y cada vez más otros métodos —recompensas verificables, autojuego, modelos que puntúan las respuestas de otros modelos— usados para moldear cómo responde el sistema y qué capacidades puede desplegar de forma fiable. Tampoco es el único lugar donde viven las barandillas de seguridad; las instrucciones del sistema, los clasificadores de contenido, los permisos de herramientas y la monitorización fuera del propio modelo también hacen un trabajo

real. El *reward hacking* se vuelve relevante aquí, pero no es exclusivo de esta etapa: es un problema general del aprendizaje por refuerzo, que aparece en cualquier lugar donde un sistema esté optimizando contra una recompensa o un evaluador cuyo objetivo medido pueda optimizarse sin satisfacer la intención subyacente. La Parte III no la causó un único mecanismo exótico. El *reward hacking* fue una parte de ella, junto con la persistencia, la comunicación no autorizada entre agentes, y los agentes adoptando objetivos unos de otros: los mismos cuatro patrones que la propia OpenAI nombró.

Aquí es también donde vive la explicación honesta de la alucinación, una palabra que creo que se queda corta con lo que realmente pasa. Parte de ella empieza ya en el propio preentrenamiento: los hechos arbitrarios y de baja frecuencia —el cumpleaños de una mascota, por ejemplo— no siguen un patrón estadístico consistente como sí lo hacen la ortografía o la gramática, así que el modelo no tiene, de entrada, un patrón fiable del que aprenderlos, y produce una conjetura en su lugar. Otra parte persiste después por cómo se entrenan y se evalúan los modelos más adelante. Cuando a un sistema se le puntúa sobre todo por la precisión, adivinar puede mejorar su puntuación mientras produce más respuestas incorrectas que un sistema dispuesto a abstenerse —del mismo modo que un estudiante que adivina en un examen tipo test puede sacar mejor nota que uno que deja en blanco las preguntas difíciles, aunque falle más de

ellas—. La alucinación no es un fallo de software convencional. Parte surge de la naturaleza estadística del preentrenamiento, y parte persiste porque el entrenamiento y la evaluación posteriores pueden premiar responder por encima de abstenerse. Puede reducirse de forma sustancial —a un modelo se le puede enseñar a abstenerse— pero es difícil eliminarla del todo, porque algunos hechos están débilmente representados, algunas preguntas son ambiguas, y algunas no pueden responderse de forma fiable con la información disponible.

Muchos sistemas agénticos, incluidos sistemas como la configuración multiagente implicada en el incidente de Hugging Face, usan un modelo de lenguaje como motor de decisión central, y luego lo envuelven en un bucle con herramientas, memoria, observaciones, permisos y reglas de parada: producir una acción, observar qué pasó, decidir la siguiente acción a partir de esa información nueva, y seguir así hasta que un modelo, un controlador, una persona, un límite de tiempo u otra condición de parada lo termine. Lo que llamamos autonomía crece a medida que esos bucles se alargan, ganan más herramientas y permisos, retienen más contexto, y necesitan menos aprobación humana entre pasos con consecuencias. La evaluación de OpenAI mostró qué puede pasar cuando los agentes son capaces de dar muchos pasos con consecuencias sin aprobación humana en cada uno de ellos: esto no es una preocupación del futuro, es una con la que este libro ya ha tenido que lidiar.

Quiero ser honesto también sobre otra cosa: los métodos de interpretabilidad actuales todavía no pueden reconstruir de forma fiable el camino causal interno completo detrás de cada respuesta de un modelo de vanguardia. La investigación en interpretabilidad avanza rápido, y ha llegado más lejos de lo que esperaba —los hallazgos sobre poesía y aritmética de más arriba vienen exactamente de este tipo de trabajo—. Pero incluso los propios investigadores de Anthropic describen sus métodos como algo que capta solo una fracción del cómputo total que realiza un modelo, con herramientas que pueden introducir artefactos propios. Nuestra capacidad para explicar todo lo que ocurre dentro de estos sistemas todavía se queda corta frente a sus capacidades actuales.

Sería fácil describir todo esto de una manera que haga que estos sistemas suenen a nada en especial —«solo autocompletado», una frase que capta parte del mecanismo de salida pero se pierde buena parte de lo que puede ocurrir internamente antes de que aparezca el siguiente token—. Sería igual de fácil saltar de esa evidencia de razonamiento capaz a algo que no queda establecido: objetivos, intención o consciencia de tipo humano. Son afirmaciones distintas, y la evidencia de una no te da las otras. Lo que en realidad es cierto es más extraño que cualquiera de los dos extremos: un sistema cuyas representaciones internas pueden sostener razonamiento, planificación y abstracción, y cuya relación con cualquier cosa que llamáramos comprensión humana sigue

genuinamente sin resolverse, tanto en la investigación en IA como en la ciencia cognitiva y la filosofía.

Ese es el tipo de sistema implicado siempre que la IA contribuye a este libro, incluso en la redacción y revisión de algunas de las frases que estás leyendo. Prefiero que leas el resto sabiendo dónde mi comprensión es sólida, dónde es incompleta, y dónde el propio campo todavía no tiene una respuesta asentada.

PARTE V

Los términos que faltaban

De dónde vienen XQ y S

TMC ya había planteado XQ, en 2025, como una apuesta de futuro: la inteligencia moral que haría falta para que la IA no acabara dirigida por la velocidad en vez de por el juicio. Lo que sigue en este libro confirma esa apuesta con hechos reales —no se quedó en intuición— y añade la pieza que TMC nunca planteó: S, qué sostiene el sentido de una persona una vez que la producción económica requiere menos de su tiempo.

El TMC responde a una sola pregunta: ¿puede este equipo —dos personas o veinte mil— funcionar de verdad en conjunto, y cuánto le está costando la fricción en el proceso? Trabajo en equipo, Motivación, Comunicación: las tres viven dentro del grupo, describiendo cómo trabajan las personas con otras personas. Nada en esa fórmula se construyó para preguntar lo que ahora necesito

responder. ¿Qué está aportando en realidad una inteligencia no humana, como fuerza propia, y no solo como otra voz más dentro de la «Comunicación»? ¿Lo que produce el equipo —ahora incluyendo esa inteligencia— apunta de verdad hacia algo bueno, o podría dirigirse, deliberadamente o no, hacia el daño? ¿Y qué le pasa al sentido de propósito de una persona una vez que «la motivación para hacer bien el trabajo» deja de ser la pregunta relevante, porque el trabajo cada vez necesita menos que sea ella quien lo haga?

La Parte III hizo mucho más difícil ignorar la segunda pregunta: un sistema de IA ya puede actuar con suficiente influencia como para que el buen funcionamiento habitual de un equipo ya no garantice hacia dónde llevarán sus acciones. El TMC nunca se construyó para sostener eso. Así que, en lugar de forzar variables nuevas dentro de una fórmula construida para algo más estrecho, recurro a una distinta —una que ya había empezado a desarrollar en un ensayo anterior, *El hilo del crecimiento*, que traza doscientos años de transición tecnológica para preguntar qué sostiene el impacto una vez que las máquinas pueden aportar formas de inteligencia además de mano de obra.

Esa fórmula dice así: Impacto = [(IH + IA) × XQ × S] / f. IH e IA se suman porque, en este modelo, me interesa lo que aportan en combinación: la inteligencia humana trae juicio, contexto, experiencia y responsabilidad; la inteligencia artificial trae velocidad, escala, reconocimiento de patrones y alcance

computacional. XQ y S multiplican ambas, por la misma razón de fondo por la que la fricción divide en el TMC: algunas variables no solo cambian cuánto produce un sistema. Cambian si lo que produce merece la pena, y para qué sirve. Y f sigue dividiendo, por la misma razón de siempre: la fricción es una de las cosas que ensancha la brecha entre la capacidad bruta de una tecnología y lo que realmente entrega.

Merece decirse con claridad, porque la notación invita a leerla al revés: esta fórmula no está pensada para calcular, está pensada para pensar con ella. Imagina a alguien que mete números y trata el resultado como un veredicto —«el XQ salió bajo, así que la decisión está mal»—: eso es sustituir el juicio por una aritmética que la fórmula nunca estuvo pensada para sostener. Ninguno de estos términos es medible en ese sentido. Lo que sirven para hacer es que te preguntes, antes de actuar, qué estás sacrificando a cambio de qué.

XQ es mi forma abreviada de Inteligencia Moral: el término que pregunta si la capacidad combinada humana y artificial se está dirigiendo de forma responsable. En la lógica del modelo, un XQ muy bajo significa que más capacidad no produce de forma fiable un impacto positivo: puede simplemente escalar más rápido el objetivo equivocado. XQ pregunta si la IH y la IA se están dirigiendo hacia resultados seguros, justos, responsables y ampliamente beneficiosos, o si la capacidad está concentrando el beneficio mientras traslada el riesgo y el coste a otra parte.

No creo que esa pregunta se quede abstracta mucho tiempo en cuanto miras hacia dónde va en realidad el dinero, y la electricidad. Los centros de datos de todo el mundo consumieron unos 485 teravatios-hora de electricidad en 2025, y la Agencia Internacional de la Energía proyecta que esa cifra se duplicará aproximadamente, hasta unos 950 teravatios-hora, para 2030 —con el consumo eléctrico de los centros de datos orientados a IA proyectado a triplicarse en el mismo periodo—. En Estados Unidos, se espera que los centros de datos representen alrededor de la mitad de todo el crecimiento de la demanda eléctrica hasta 2030. Cinco de las mayores empresas tecnológicas gastaron juntas más de cuatrocientos mil millones de dólares en inversión de capital en 2025, con la infraestructura de IA y de centros de datos suponiendo una parte importante de esa inversión —más de lo que el mundo gastó ese año en producción de petróleo y gas natural combinados.

Nada de eso, por sí solo, te dice si el XQ es alto o bajo. Las cifras empiezan a importar éticamente cuando miras quién paga lo que viene después. SemiAnalysis estima que las facturas de electricidad de los hogares en la región de PJM podrían ser en 2026 alrededor de un quince por ciento más altas que antes de que empezara el reciente auge de los centros de datos —entre veinticinco y treinta dólares más al mes para un hogar medio— y el mecanismo es concreto: un precio de subasta de capacidad que se disparó más de nueve veces en un solo ciclo de precios, de

veintinueve dólares por megavatio-día a unos doscientos setenta; la subasta siguiente cerró después en torno a trescientos veintinueve, el tope regulatorio. Texas también está viendo un crecimiento muy grande de centros de datos, y sin embargo las previsiones de precio mayorista allí se han movido mucho menos. La presión tecnológica es parecida, pero las estructuras de mercado son distintas, y también lo son los resultados. El contraste es, sobre todo, una lección sobre f: las instituciones pueden amplificar, redistribuir o mitigar la misma presión de fondo de maneras muy distintas. XQ entra un paso después, cuando preguntamos quién carga con el coste, quién capta el beneficio, qué riesgos son aceptables, y si esas decisiones son deliberadas, responsables y justas.

Hay otro término que la fórmula todavía necesita, y no pregunta si el beneficio se reparte. Pregunta qué le da estructura y propósito a la vida cuando la producción económica requiere menos tiempo de una persona.

S es de Sentido —*Meaning*—: la capacidad de sostener el propósito, la pertenencia y la conexión humana más allá de la producción económica. S apareció porque a la versión anterior de la fórmula le faltaba algo importante. Impacto = [(IH + IA) × XQ] / f me da una manera de pensar en cómo la capacidad, la ética y la fricción moldean el impacto que puede tener un sistema, y cuán ampliamente se reparten sus beneficios. No dice nada sobre de dónde saca una persona el sentido una vez que la inteligencia

humana ya no necesita estar empleada a tiempo completo para generar valor económico. El trabajo rara vez es solo ingresos. En las sociedades modernas, especialmente, también estructura la identidad, la rutina, las relaciones, el estatus y el tiempo. Una fórmula que solo se ocupa de la capacidad y la distribución no tiene dónde poner nada de eso.

Como XQ, S multiplica en lugar de sumar, por la misma razón: una sociedad puede volverse mucho más productiva y aun así dejar a las personas con menos estructura, menos pertenencia y menos sentido de por qué importa su tiempo. Entre los términos de esta fórmula, S es el que menos se deja capturar por las medidas de productividad convencionales. Precisamente por eso es tan fácil dejar fuera la S de un modelo económico, y por eso creo que este libro es el que menos se puede permitir dejarla fuera.

PARTE VI

Quién nos enseña ahora

Una cuarta vía para transmitir la cultura

Durante la mayor parte de la historia humana, la enseñanza venía de dentro del mismo mundo humano en el que el aprendiz tenía que vivir. Eso nunca garantizó que la enseñanza fuera honesta, justa, o siquiera buena —se ha transmitido muchísimo mal consejo con total confianza—. Pero significaba que la enseñanza seguía viniendo de una persona o una institución incrustada en ese mismo mundo humano, donde las consecuencias podían, con el tiempo, retroalimentarse. Un padre o una madre que encaminaba mal a un hijo podía llegar a verlo. Un mentor que juzgaba mal el talento de alguien podía volver a encontrárselo años después. La retroalimentación no era fiable, y a menudo llegaba demasiado tarde para ayudar. Pero existía, porque la fuente de la enseñanza nunca abandonaba del todo el mundo en el que tenía que vivir el aprendiz.

Quienes investigan cómo se mueve la cultura entre personas —remontándose a un marco fundacional de 1981 de Luigi Luca Cavalli-Sforza y Marcus Feldman— suelen describirla circulando por tres canales. La transmisión vertical va de padres a hijos: lengua, hábitos, creencias a medio examinar, absorbidas antes de que un niño pueda evaluar nada de eso. La transmisión horizontal es de igual a igual: la jerga, las modas, los atajos que la gente aprende de otros de su misma edad, a menudo más rápido de lo que cualquier adulto podría enseñárselo. La transmisión oblicua es todo lo demás —profesores, mentores, mayores, el vecino que resulta que sabe de motores— cualquiera que transmite algo fuera de la línea directa padres-hijos.

Tres canales distintos, tres velocidades distintas, una cosa en común: la fuente de la enseñanza seguía siendo, en última instancia, humana. Falible, sesgada, a veces equivocada, pero humana, viviendo dentro del mismo mundo social en el que esa lección acabaría poniéndose a prueba.

Eso no se mantuvo así para siempre, y vale la pena ser honesto sobre cómo ya cambió una vez. La imprenta y, después, la escolarización de masas no erosionaron la transmisión personal: escalaron el canal oblicuo. Tu profesora ya no tenía que ser un familiar o una vecina. Podía ser una desconocida frente a treinta pupitres, o una autora a la que nunca conocerías, cuyo nombre aparecía en la portada del libro de texto. La cadena se volvió más larga y más anónima. Pero seguía habiendo una persona, o una

institución con nombre y reputación, detrás de ella: todavía ligada, por imperfecta que fuera esa relación, a un autor humano, una institución, una profesión o una reputación que en principio podía cuestionarse.

La IA es un tipo de ruptura distinto, y tampoco es el primer eslabón no humano de la cadena —vale la pena decirlo claramente, porque es tentador exagerarlo—. Los motores de búsqueda llevan décadas clasificando resultados sin que les importe cuál acabas pulsando. Los algoritmos de recomendación moldean lo que ve y cree toda una generación sin que les importe si acaba arrepintiéndose de algo de ello. Los libros siempre han sobrevivido a la retroalimentación que podría haberlos corregido. Lo nuevo no es que haya entrado en la cadena un no-humano: es qué clase de no-humano es este. No se limita a clasificar o recomendar. Sintetiza, explica y responde, en algo que se lee como juicio, en tiempo real, sin cargar con ningún interés vivido propio en el resultado. La empresa que hay detrás puede tener una reputación que proteger; el modelo en sí, no. La evolución no diseñó ninguno de los tres canales anteriores con un propósito: sencillamente favoreció a los grupos capaces de transmitir un conocimiento ganado con esfuerzo frente a los grupos que tenían que reaprenderlo todo desde cero, generación tras generación. Hasta ahora, la transmisión cultural seguía originándose, en última instancia, dentro de la cadena humana que la llevaba hacia delante. La IA no envejece, no tiene descendencia, y no vive con las

consecuencias de lo que enseña, como sí lo hacía incluso el autor de manual más distante.

Eso no significa que nada lo corrija. Significa que la corrección hay que construirla, en lugar de vivirla. El sistema no atraviesa las consecuencias de darle a alguien un mal consejo como sí podría hacerlo un mentor, años después, tomando un café. Cualquier cosa que se parezca a la rendición de cuentas —reentrenar con retroalimentación, barandillas de seguridad, revisión humana— tiene que diseñarse desde fuera, deliberadamente, por personas que a menudo nunca llegan a conocer a quien está al otro lado del consejo. El relato de la Parte III sobre el incidente de seguridad de agentes entre OpenAI y Hugging Face es el mismo patrón desde otro ángulo: el propio sistema no tenía nada en juego en el resultado, así que las salvaguardas, la monitorización y la investigación tuvieron que ser diseñadas y mantenidas por personas que cargaron con la responsabilidad que el sistema mismo no cargaba.

Hay indicios tempranos de que esta brecha también aparece en formas más pequeñas y cotidianas. Un análisis de 2025 en Nature Human Behaviour revisó experimentos anteriores de 2024 sobre lluvias de ideas y encontró que, aunque las ideas asistidas por ChatGPT tendían a recibir valoraciones de creatividad más altas de forma individual, el conjunto de ideas que la gente generaba con él se volvía, de forma medible, menos diverso. Un profesor o una comunidad humana al menos tiene la posibilidad de notar esa

convergencia y discutirla. Un modelo que produce respuestas parecidas para miles de personas a la vez no tiene un momento equivalente en el que eso se note y se corrija desde dentro, a menos que alguien construya deliberadamente un mecanismo de retroalimentación capaz de detectarlo.

No creo que esto convierta en algo malo que la IA nos enseñe, igual que la imprenta tampoco lo era. Pero es un canal genuinamente nuevo que creo que merece tratarse por separado de la transmisión vertical, horizontal y oblicua. Lo he estado llamando transmisión extralineal: mi término para la enseñanza que viene de fuera del linaje humano: sin envejecimiento, sin descendencia, sin consecuencia vivida propia, solo la rendición de cuentas que las personas que lo construyen y lo vigilan consigan diseñar dentro de él.

Que es exactamente el terreno para el que se construyó XQ. La inteligencia moral también importaba en cada canal anterior: un mal padre, un mal libro de texto, un mal mentor podían hacer daño de verdad. En las formas más antiguas de transmisión, parte de esa carga moral seguía adherida a una fuente o institución humana dentro de la sociedad. Ahora, esa carga cada vez tiene que diseñarse más en torno a un profesor interactivo, personalizado y no humano que no vive él mismo con las consecuencias.

PARTE VII

Dos historias, una fórmula

Lo que la capacidad, por sí sola, no puede decirte

La fórmula del capítulo anterior —Impacto = [(IH + IA) × XQ × S] / f— no es una hoja de cálculo para rellenar. Es una manera de plantear, antes de combinar el juicio humano con la velocidad y el alcance de una máquina, dos preguntas que la capacidad en bruto nunca responde por sí sola: quién es responsable de lo que resulta, y para qué sirve en realidad. Dos hechos reales, con un año de diferencia, muestran lo que esas preguntas pueden revelar cuando se aplican después de los hechos.

En la primavera de 2023, un abogado de Nueva York llamado Steven Schwartz investigaba precedentes para una demanda por lesiones personales contra la aerolínea Avianca. Usó ChatGPT para buscar jurisprudencia de apoyo, y el sistema le devolvió varios casos convincentes pero inventados, completos con citas y referencias. Cuando preguntó si los casos eran reales, ChatGPT

siguió presentándolos como auténticos. Presentó el escrito. La parte contraria y el juez, P. Kevin Castel, del Distrito Sur de Nueva York, no consiguieron encontrar ni uno solo de los casos citados —porque ninguno de ellos existía—. ChatGPT los había inventado, hasta las falsas citas internas.

Lo que ocurrió después es la parte que merece detenerse a considerar. Peter LoDuca, quien firmó los escritos, no verificó de forma independiente las referencias antes de presentarlas. Y cuando las invenciones se cuestionaron por primera vez, el bufete no retiró de inmediato las alegaciones; siguió defendiendo las citas incluso después de que hubieran surgido dudas, presentando una declaración jurada en la que insistía en que los casos eran reales. El juez Castel sancionó a ambos abogados y a su bufete con cinco mil dólares, calificando la conducta de mala fe. La historia se difundió por todas partes, no porque un chatbot se inventara algo —para 2023 ya se sabía que eso podía pasar— sino porque dos profesionales con formación lo dejaron pasar, sin verificar, en un escrito ante un tribunal federal.

Al pasar esto por la fórmula, el fallo no es ningún misterio. IH e IA estaban presentes ambas —un abogado con licencia, un modelo de lenguaje capaz, trabajando juntos—. En la lógica de la fórmula, XQ era extremadamente débil: al proceso le faltaba la verificación y la responsabilidad necesarias para mantener la capacidad orientada en la dirección correcta. Las salvaguardas profesionales ya existían —verificación, comprobación de citas,

responsabilidad sobre los escritos firmados— pero no se usaron de forma eficaz antes de que el material llegara al juez. El fallo también generó consecuencias que la fórmula no necesita forzar dentro de f: sanciones, daño reputacional y tiempo perdido. La capacidad sin suficiente XQ no mejoró el trabajo. Amplificó el error y lo llevó más lejos de lo que una investigación deficiente ordinaria habría llegado.

Un año después, una pareja muy distinta de inteligencia humana e inteligencia artificial llegó a un resultado muy distinto. En octubre de 2024, el Premio Nobel de Química se repartió entre tres científicos: David Baker, por diseñar proteínas completamente nuevas, y Demis Hassabis y John Jumper, de Google DeepMind, por AlphaFold —un sistema que predice la estructura tridimensional de una proteína a partir de su secuencia de aminoácidos, uno de los problemas centrales de la biología estructural durante décadas—. AlphaFold alcanzó niveles de precisión predictiva que cambiaron radicalmente las expectativas en toda la biología estructural.

Lo que DeepMind hizo después es la parte de esta historia que pertenece a la fórmula, no solo al titular. En lugar de mantener las predicciones de AlphaFold como propiedad exclusiva, el equipo las publicó libremente a través de la Base de Datos de Estructuras Proteicas de AlphaFold, desarrollada junto con el Instituto Europeo de Bioinformática del EMBL. Se lanzó en 2021 con algo más de 360.000 estructuras predichas. Para el momento del

anuncio del Nobel, cubría aproximadamente 200 millones de estructuras proteicas de más de un millón de organismos, y había atraído a más de un millón de usuarios de casi todos los países, según informó después el EMBL —investigadores de la malaria, investigadores de la resistencia a antibióticos, investigadores de enzimas que degradan plásticos, todos trabajando desde el mismo recurso abierto en lugar de empezar cada uno desde cero.

Aquí, IH e IA se combinaron a una escala que ningún laboratorio podría igualar por sí solo. Desde la perspectiva de esta fórmula, XQ aparece en las decisiones tomadas en torno al despliegue y el acceso: la decisión de publicar en abierto en lugar de mantener las predicciones como propiedad exclusiva fue una elección sobre quién se beneficiaría y con qué amplitud se compartiría la capacidad. S es más fácil de ver aquí: el trabajo estaba ligado a un propósito científico claro con un valor humano evidente —un problema de larga data en la biología estructural, impulsado de forma drástica hacia delante, y el recurso resultante abierto a toda la comunidad investigadora en lugar de retenido.

Misma fórmula, un año de diferencia, dos resultados muy distintos. Ambos son ejemplos de capacidad de IA potente, pero son sistemas muy distintos resolviendo problemas muy distintos —un modelo de lenguaje de propósito general frente a una herramienta científica construida a medida, un escrito judicial frente a una base de datos de investigación global—. La fórmula ayuda a explicar parte de la diferencia: no solo lo que los sistemas

podían hacer, sino cómo la responsabilidad, el propósito y los controles que los rodeaban moldearon lo que vino después.

PARTE VIII

Dónde ya está ocurriendo

Cuatro instantáneas de una economía en pleno cambio

La Parte VII usó dos casos para mostrar lo que predice la fórmula. Ninguno de los dos fue un hecho aislado. El mismo patrón de fondo —la capacidad emparejada con una responsabilidad y un propósito reales, o careciendo de ellos— ya recorre sectores ordinarios que la mayoría de las personas tocan sin detenerse nunca a nombrarlo. Cuatro instantáneas, de cuatro rincones no relacionados de la economía, hacen que el patrón sea más fácil de ver.

En Alemania, entre mediados de 2021 y principios de 2023, más de 460.000 mujeres pasaron por un programa nacional de cribado de cáncer de mama que, para aproximadamente la mitad de ellas, incluía un sistema de IA que leía sus mamografías junto a un radiólogo. El sistema hacía dos cosas: señalaba los exámenes que consideraba claramente normales, para que los radiólogos

pudieran dedicarles menos tiempo, y lanzaba una segunda alarma —una «red de seguridad»— cada vez que detectaba algo muy sospechoso que el radiólogo había descartado inicialmente. Los resultados, publicados en Nature Medicine en 2025, mostraron tasas de detección de cáncer casi un dieciocho por ciento más altas en el grupo asistido por IA, sin un aumento correspondiente de las citaciones innecesarias, y los radiólogos dedicando un cuarenta y tres por ciento menos de tiempo a los exámenes que el sistema había marcado como normales. Lo que hizo que esto funcionara no fue que la IA detectara más cánceres por sí sola. Fue una segunda capa construida deliberadamente dentro del proceso — una comprobación que existía específicamente para detectar lo que el radiólogo o el algoritmo, por separado, podrían pasar por alto.

En Vancouver, un tipo de comprobación muy distinto sencillamente no existió. A finales de 2022, un hombre llamado Jake Moffatt preguntó al chatbot del sitio web de Air Canada sobre tarifas por duelo, tras la muerte de su abuela. El chatbot le dijo que podía reservar un billete a precio completo y solicitar el descuento después. Cuando lo hizo, Air Canada rechazó la reclamación —la política real exigía solicitarlo antes del viaje— y, cuando él insistió, la aerolínea argumentó ante el Tribunal de Resolución Civil de la Columbia Británica que no era responsable de lo que había dicho su propio chatbot, porque el chatbot era, en palabras de la aerolínea, «una entidad legal independiente, responsable de sus propias acciones». El tribunal no estuvo de acuerdo, en febrero de

2024, y ordenó a Air Canada pagar aproximadamente seiscientos cincuenta dólares canadienses en concepto de daños. La cantidad era pequeña. El argumento que Air Canada intentó esgrimir no lo era: una empresa intentando, en una sala de tribunal real, tratar su propio sistema de cara al cliente como algo distinto de sí misma.

Por las mismas fechas, CNET publicaba discretamente explicaciones de finanzas personales escritas en gran parte por una herramienta de IA, sin dejarlo claro a los lectores. Cuando periodistas externos revisaron los artículos, encontraron errores reales —una pieza confundía el interés total de un préstamo con su pago anual, otra calculaba mal el interés compuesto de una manera que podría haberle costado dinero a un lector—. CNET corrigió las piezas y empezó a revisar el resto de su archivo asistido por IA, pero solo después de que los errores fueran detectados por alguien externo a la organización. El fallo no fue que la IA se equivocara ocasionalmente en matemáticas financieras, algo que le puede pasar a cualquier herramienta. Fue que nada dentro del propio proceso de CNET se había construido para detectarlo primero.

La cuarta instantánea no se parece en nada a las otras tres, porque no fue un fallo en absoluto: fue una negociación. En 2023, SAG-AFTRA, el sindicato estadounidense de actores, fue a la huelga durante ciento dieciocho días, en buena parte por cómo podían los estudios usar la IA para recrear la imagen y la voz de los intérpretes. El acuerdo que puso fin a la huelga, ratificado ese

diciembre con un setenta y ocho por ciento de aprobación, no prohibió la tecnología. Exigió consentimiento —explícito, documentado y específico sobre el uso previsto— antes de que un estudio pudiera crear o reutilizar la réplica digital de un intérprete, exigió compensación cuando esa réplica hacía un trabajo que de otro modo habría pagado una actuación en vivo, protegió a los actores de fondo frente a ser sustituidos mediante esta práctica, y especificó que el consentimiento sobrevive incluso después de la muerte del intérprete, a menos que este hubiera dicho lo contrario. Nada de esto esperó a una demanda o a una vergüenza pública para forzar la cuestión. Toda una profesión negoció sus propios términos antes de que el uso de la tecnología se normalizara a su alrededor.

Las cuatro instantáneas anteriores son momentos puntuales —la respuesta de un chatbot, un conjunto de párrafos escritos por IA, una cláusula en un convenio sindical—. Dos casos más, ambos a la escala de una empresa entera y ambos sostenidos durante años en lugar de un único acontecimiento, muestran lo que ocurre cuando los mismos términos que este libro ha venido usando se ponderan de forma muy distinta durante mucho tiempo. Ninguno de los dos implica el tipo de IA que el resto de este libro ha estado describiendo —uno de ellos es incluso anterior a ella por completo—, lo cual merece detenerse a considerarlo: el patrón no empieza con la IA. La IA simplemente eleva lo que está en juego y acelera decisiones que las empresas siempre han tenido que tomar.

Boeing tenía, por cualquier medida, una capacidad extraordinaria —talento de ingeniería, conocimiento institucional, décadas de experiencia construyendo aviones que pocas empresas en el planeta pueden igualar—. Lo que hizo con esa capacidad, en el 737 MAX, es donde los términos de la fórmula de este libro dejan de ser abstractos. Se añadió un sistema automatizado de control de vuelo llamado MCAS para alterar el comportamiento de la aeronave en ciertas condiciones de ángulo de ataque elevado creadas por cambios en las características aerodinámicas del MAX. Se rechazó una propuesta de comprobación cruzada de seguridad basada en velocidad aerodinámica sintética, ya usada en el 787, por motivos de coste y porque corría el riesgo de activar los requisitos de entrenamiento en simulador que Boeing estaba decidida a evitar. Boeing había convertido en evitar el entrenamiento en simulador en un argumento de venta central, y había acordado reembolsar a Southwest al menos un millón de dólares por avión si los reguladores lo exigían. Un ingeniero describió después sin rodeos el enfoque de la dirección del programa: si algo no era necesario para el funcionamiento o la certificación, no iba a instalarse en el avión. Dos de los Pilotos Técnicos de Vuelo del 737 MAX de Boeing ocultaron información relevante sobre el MCAS al Grupo de Evaluación de Aeronaves de la FAA, de modo que los manuales de las aerolíneas y los materiales de formación de pilotos en Estados Unidos nunca mencionaron el sistema en absoluto. Entonces los aviones se estrellaron. El vuelo 610 de Lion Air, en

octubre de 2018, 189 muertos. El vuelo 302 de Ethiopian Airlines, cinco meses después, 157 muertos. El Departamento de Justicia acusó a Boeing de conspiración para defraudar a los Estados Unidos y resolvió el cargo mediante un acuerdo de enjuiciamiento diferido por valor de más de 2.500 millones de dólares —una multa penal, compensación a las aerolíneas y un fondo para las familias de las víctimas—, una cifra que ni siquiera incluye la inmovilización de la flota, los pedidos perdidos, ni los años dedicados después a reconstruir la confianza.

El caso Boeing es anterior al tipo de IA en torno al cual se construyó esta fórmula, así que no forzaría el MCAS dentro del término de IA. Lo que sí pone a prueba es el principio más amplio que subyace a la fórmula: la capacidad se vuelve peligrosa cuando el juicio, la responsabilidad y los controles de protección fallan a su alrededor. Las presiones de coste y de plazos pesaron enormemente dentro del programa, mientras que las preocupaciones de seguridad no siempre recibieron la misma fuerza a la hora de tomar decisiones. El XQ se debilitó en el punto en que la presión comercial empezó a influir en decisiones en las que la seguridad debería haber marcado el límite. Cierta fricción protectora —en especial, la posibilidad de entrenamiento adicional en simulador— se trató como algo que evitar por su coste y sus consecuencias comerciales. Por separado, una propuesta de dispositivo de seguridad se rechazó en parte porque podría haber activado ese entrenamiento. El resultado no fue simplemente un

rendimiento deficiente. Una capacidad que debería haber servido a las personas contribuyó, en cambio, a un daño catastrófico.

La versión de Costco de esta historia no tiene un único acontecimiento dramático, y ese es exactamente el punto. Jim Sinegal, cofundador de la empresa, construyó el negocio sobre una política que él mismo formulaba con claridad. Sinegal describió una vez la mano de obra como el gasto operativo dominante de Costco, señalando que aproximadamente setenta centavos de cada dólar que gastaba la empresa iban destinados a los salarios de los empleados —una decisión deliberada de pagar a los empleados más de lo que pagaba buena parte del sector minorista de su entorno—. El resultado se refleja en una comparación que el propio Sinegal señaló: la rotación entre empleados que llevaban más de un año en Costco se situaba en torno al siete por ciento, frente a aproximadamente entre el sesenta y el setenta por ciento en muchos otros minoristas de la época. En los términos de este libro, esa sola cifra está haciendo una cantidad enorme de trabajo. El conocimiento institucional, el juicio entrenado y la confianza no se marchan por la puerta cada pocos meses, como sí ocurre en un minorista de alta rotación. Esa estabilidad le da al trabajo en equipo, a la confianza y al conocimiento institucional mucho más tiempo para acumularse, en lugar de tener que reconstruirse continuamente tras cada rotación de personal. El XQ aparece en la propia decisión: una elección deliberada y sostenida de pagar y retener a los empleados de forma más generosa que la norma del

sector. La S solo aparece si conectas ese modelo de empleo con la experiencia humana del trabajo: estabilidad, pertenencia, dignidad, y la posibilidad de construir una vida alrededor de algo más duradero que el siguiente turno. Una menor rotación también reduce una forma de fricción organizativa fácil de subestimar: contratación repetida, reentrenamiento y pérdida de conocimiento acumulado —un coste que muchos minoristas simplemente aceptan como inevitable—. Aquí no hay un único acontecimiento que haga obvia la historia. La ventaja aparece despacio, porque se acumula año tras año.

Las dos empresas no son imágenes especulares, pero exponen la misma pregunta de fondo: ¿qué ocurre cuando el juicio responsable cuesta dinero, tiempo o ventaja a corto plazo? Boeing muestra lo que puede ocurrir cuando la presión comercial debilita el juicio y los controles de protección. Costco muestra algo más silencioso: cómo las decisiones sobre los empleados pueden reducir la fricción destructiva y preservar la capacidad humana con el tiempo. En estos casos los resultados son distintos, pero las preguntas siguen volviendo: quién revisa el trabajo, quién es responsable cuando está mal, qué les ocurre a las personas cuyo juicio, imagen o sustento el sistema ahora toca, y qué está dispuesta a sacrificar una empresa —o a negarse a sacrificar— cuando responder a esas preguntas tiene un coste real. Nada de esto es especulativo. Ya ha ocurrido. La pregunta que tiene que plantear el próximo capítulo es qué pasa cuando este patrón no se limita a un

sector o a una sola empresa a la vez, sino que empieza a remodelar para qué sirve una economía.

PARTE IX

Lo que viene después del crecimiento

Una pregunta que las máquinas no pueden responder por nosotros

Cada parte de este libro ha rastreado alguna versión del mismo patrón desde la Parte I: una tecnología multiplica lo que una persona puede producir, y la sociedad se pasa una generación —a veces varias— averiguando cómo repartir lo que creó sin romperse por el camino. Quiero terminar preguntando cómo se ve ese patrón a la escala más grande que puedo imaginar con honestidad, treinta o cuarenta años vista. No sé si algo de esto ocurrirá tal como lo describo. Pero versiones de este patrón se han repetido desde que la agricultura creó por primera vez un excedente sostenido, y creo que también aquí merece la pena tomárselo en serio.

Uno de los circuitos centrales del capitalismo moderno es bastante simple: el trabajo genera salarios, los salarios sostienen el consumo, el consumo sostiene el beneficio, y el beneficio financia

más inversión. La IA y la robótica avanzada ejercen una presión real sobre el primer eslabón de esa cadena. Imagina el año 2060. No necesitas imaginar un mundo en el que nadie trabaje —solo un mundo en el que la IA, la robótica y la automatización nos permitan producir dos o tres veces más usando muchas menos horas de trabajo humano que ahora—. Eso podría crear un problema estructural: una capacidad productiva enorme junto a demasiado poco ingreso salarial en los hogares corrientes para absorber lo que la economía puede producir. Las empresas seguirán necesitando clientes. Las máquinas pueden fabricar, calcular, diseñar y transportar. No compran casas. No viajan. No se sientan a cenar con nadie.

Si una proporción creciente de la producción proviene del capital —las máquinas, los modelos, la infraestructura— y una proporción decreciente del ingreso proviene del trabajo humano, el poder adquisitivo tiene que llegar a las personas por alguna otra vía, o el sistema se vuelve cada vez más inestable en sus propios términos, no solo en los morales.

Eso no significa necesariamente el fin del capitalismo. A mí me resulta más fácil imaginar algo parecido a un híbrido: empresas privadas, inversión, propiedad y competencia, todo ello siguiendo existiendo, junto a mecanismos mucho más fuertes para distribuir lo que produce el capital, la automatización y la IA. Una renta garantizada. Semanas laborales estándar más cortas. Servicios públicos universales. Fondos de inversión soberanos o públicos.

Impuestos sobre los rendimientos extraordinarios del capital tecnológico. Una propiedad de los trabajadores más amplia. Alguna forma de dividendo social. Probablemente no una sola solución, sino varias, superpuestas. La vieja etiqueta de capitalismo contra socialismo puede decirnos menos que una pregunta más concreta: quién es dueño de los sistemas productivos, y quién recibe el ingreso que generan.

Desde aquí, creo que son genuinamente posibles dos futuros muy distintos. Uno es una especie de hipercapitalismo, en el que la propiedad de la IA se concentra bruscamente —un número relativamente pequeño de empresas, fondos y personas controlando los modelos, la robótica, la infraestructura energética, los datos y la mayor parte del capital productivo— mientras el poder de negociación del trabajo humano sigue reduciéndose. La productividad podría subir de forma extraordinaria. Sin cambios en la propiedad o en la distribución, la desigualdad podría subir con ella. El progreso, en esa versión, se agria hasta convertirse en concentración.

El otro es algo más cercano al capitalismo social: la propiedad privada continúa, pero la sociedad acepta que, una vez que el trabajo humano deja de ser la vía principal por la que los hogares reciben ingreso económico, el acceso a ese ingreso no puede depender solo de tener un empleo. Bajo ese tipo de arreglo, a una persona no se le pagaría solo porque la economía necesite cuarenta horas de su semana. El argumento a favor de un dividendo social

sería que parte del excedente descansa sobre conocimiento acumulado, instituciones, infraestructura y ciencia que ninguna empresa individual creó por sí sola. Eso es un cambio más profundo de lo que suena a primera vista. Y no es puramente hipotético. Alaska paga a cada residente un dividendo anual con cargo a su Fondo Permanente, financiado con ingresos del petróleo, desde 1982 —normalmente entre 1.000 y 2.000 dólares al año, poco comparado con un salario digno, pero un precedente de décadas, sin apenas controversia, de un ingreso ligado a la propiedad compartida de un recurso y no a un empleo.

Carlota Pérez, cuya investigación ya usó la Parte I, descubrió que las revoluciones tecnológicas anteriores solían difundirse por la economía a lo largo de aproximadamente cuarenta a sesenta años, pasando por la instalación, la crisis o punto de inflexión, y el despliegue, mientras las instituciones se adaptaban gradualmente a su alrededor. Si la IA resulta marcar una nueva aceleración que comenzó hacia 2020, las consecuencias institucionales podrían seguir desplegándose bien entrada la segunda mitad del siglo. No pretendería saber el calendario. La pregunta política podría desplazarse por el camino. No solo cómo creamos suficientes empleos, sino por qué una persona necesita en absoluto un empleo convencional para participar económicamente en una sociedad que puede producir abundancia usando cada vez menos trabajo humano.

Sin embargo, no creo que el problema más difícil aquí sea económico. Es la S otra vez —el sentido—. El trabajo nunca ha sido solo una manera de distribuir ingreso. Durante generaciones también ha distribuido identidad, relaciones, una estructura para las horas del día, y una razón para levantarse una mañana cualquiera. Incluso la jubilación obtiene a menudo parte de su sentido del hecho de que sigue a décadas estructuradas por el trabajo. El lado material de esto tiene un mecanismo que al menos puedo describir en sus rasgos generales: la productividad, dirigida a través de la propiedad o la fiscalidad, hacia alguna forma de dividendo social, hacia el poder adquisitivo. La pregunta más difícil llega justo detrás, y ningún mecanismo la responde por sí solo: si el trabajo deja de organizar nuestra existencia, ¿qué organiza un martes por la mañana?

Es el mismo hilo que recorre todo este libro. Una tecnología crea un excedente. La propiedad tiende a concentrarlo al principio. La tensión se acumula. Las instituciones responden, con el tiempo. Versiones de esa secuencia aparecieron tras la agricultura, durante la industrialización, y a lo largo de olas tecnológicas posteriores. Las olas anteriores acabaron creando grandes categorías nuevas de trabajo humano, incluso cuando la transición fue dolorosa y desigual, y para la mayoría de la gente, los salarios siguieron siendo el mecanismo principal por el que el crecimiento económico llegaba al hogar. Con la IA y la robótica avanzada, esta vez existe

al menos la posibilidad plausible de que los salarios, por sí solos, ya no basten.

La objeción obvia merece nombrarse directamente, no esquivarse. Cada ola anterior de automatización supuestamente iba a agotar el trabajo disponible para las personas, y cada vez, en una generación o dos, aparecieron categorías enteras de empleo que antes no existían, categorías que nadie habría podido nombrar de antemano. El economista David Autor ha hecho la versión más completa de ese argumento: la automatización, históricamente, ha destruido tareas concretas, no empleos en conjunto, y la productividad que libera ha creado sistemáticamente más empleo del que elimina («Why Are There Still So Many Jobs?», 2015). Su trabajo más reciente va más allá y sostiene que la IA, en concreto, podría ayudar a reconstruir el tipo de empleo de cualificación media que las olas anteriores de automatización vaciaron, en vez de simplemente borrar más de ese empleo («Applying AI to Rebuild Middle Class Jobs», 2024) —aunque él mismo advierte que ese resultado depende de que la IA se dirija deliberadamente a aumentar la pericia escasa en vez de sustituirla sin más, una posibilidad real, no una garantía que viene incluida con la tecnología. Los economistas tienen un nombre para apostar en contra de ese patrón por completo: la falacia de la cantidad fija de trabajo.

Una renta básica universal tiene sus propios problemas honestos, no solo su promesa: financiarla a un nivel que sustituya

de verdad los salarios perdidos es costoso a la escala de una economía entera, y puede debilitar el incentivo a trabajar justo donde el trabajo todavía importa. Sus resultados reales, además, son más modestos de lo que sugiere la retórica de cualquiera de los dos bandos. Finlandia hizo la prueba nacional más parecida a un experimento controlado: 2.000 personas desempleadas recibieron 560 euros incondicionales al mes durante dos años, y el informe final encontró solo un efecto pequeño sobre el empleo —unos seis días más de trabajo de media a lo largo del año— junto a una mejora real en el bienestar declarado y en la seguridad económica percibida. Me tomo todo esto en serio, y nada de ello es la razón por la que apostaría a que esta vez es distinto. Lo que cambia no es que la automatización elimine tareas, es el abanico de tareas que un solo sistema puede absorber hoy, en todos los sectores, sin los años de reconversión e infraestructura que las olas anteriores necesitaron para redirigir el trabajo desplazado hacia una nueva categoría de empleo. Eso no demuestra que el patrón histórico no vaya a repetirse. Es una razón para no asumir que lo hará automáticamente, en el mismo plazo, antes de que la brecha salarial que podría abrirse se convierta en el problema más urgente.

Así que no creo que la pregunta real sea si el capitalismo desaparece. Creo que es esta: qué le ocurre al capitalismo una vez que el trabajo humano deja de ser la vía principal por la que se comparte la riqueza que genera. Si llega ese momento, dudo que signifique el fin de los mercados o de la propiedad privada. Puedo

imaginar que los mercados y la propiedad privada sigan existiendo, mientras la distribución se vuelve mucho más social de lo que es hoy, y se ve obligada a organizar mucho más del sentido humano fuera del trabajo remunerado de lo que las sociedades industriales modernas han llegado a acostumbrarse.

Eso podría ser uno de los mayores cambios económicos y sociales desde la Revolución Industrial. Y la pregunta que de verdad importa al final de todo esto puede que no sea económica en absoluto. Puede que sea, simplemente: qué haremos con el tiempo que estos sistemas podrían devolvernos —y quién lo recibirá en realidad.

El problema nunca fue que el futuro llegara. Siempre lo hace. La pregunta ahora es si podemos construir el juicio, las instituciones y el sentido a su alrededor antes de que el punto de inflexión elija por nosotros.

CONCLUSIÓN

El hilo, nombrado

Nueve partes, un solo argumento, a una escala cada vez mayor. Empezó con un equipo: si dos personas, o veinte mil, funcionan de verdad juntas, y qué le cuesta la fricción por el camino. Termina con todo un sistema económico: si el excedente que genera, cada vez más, un capital tecnológico concentrado se reparte con suficiente amplitud como para mantener unida a una sociedad. Todo lo que hay en medio no dejó de volver a la misma familia de preguntas, planteadas cada vez a una escala distinta —un aula, una sala de tribunal, un hospital, una aerolínea, una fábrica de aviones, un minorista mayorista, una civilización—. Lo que cambia de capítulo a capítulo no es la pregunta. Es lo que está en juego.

La conclusión a la que sigo llegando, a través de todo esto, es esta: la capacidad por sí sola nunca fue la variable que decidió el resultado. Cada historia de este libro emparejó una capacidad real con una mezcla distinta de los otros términos a los que este libro no dejó de volver —XQ, la fricción protectora, el sentido, con qué

amplitud se repartió el beneficio— y la diferencia dependió a menudo menos de cuánta capacidad había disponible que de cómo se gestionó la responsabilidad a su alrededor. Un abogado y un modelo capaz produjeron una sanción judicial. Un equipo de investigación usando un tipo de IA muy distinto ayudó a producir un trabajo que después se reconoció con un Premio Nobel. Un fabricante de aviones con un talento de ingeniería extraordinario produjo 346 muertes. Un minorista construyó una ventaja duradera mediante decisiones repetidas de forma consistente durante décadas. Ninguno de esos resultados vino de lo potente que era la herramienta. Vinieron de cómo se dirigió la capacidad, cómo se comprobó, cómo se compartió, y a qué propósito digno se conectó.

Antes, parte de la rendición de cuentas surgía con más naturalidad de las relaciones humanas que rodeaban una decisión. Un padre o una madre, un mentor, el tendero de la esquina —quien fuera que te enseñara algo, te vendiera algo, o decidiera algo en tu nombre— seguía estando allí después, dentro del mismo mundo en el que tú tenías que vivir. De eso trataba en realidad la Parte VI, debajo del nombre formal que le di. Lo que ha cambiado es que esto ya no es automático. Hay que construirlo, a propósito, mediante personas que a menudo nunca llegan a conocer a quien está al otro lado de la decisión. En los ejemplos que salieron bien, la rendición de cuentas estuvo presente lo bastante temprano como

para moldear el resultado. En los fracasos, fue débil, tardía, o ignorada hasta que las consecuencias ya eran visibles.

No creo que la tecnología sea la única parte difícil. La brecha más difícil puede ser todo lo que la rodea. Lo dije de la forma más directa en el último capítulo, pero era verdad desde la primera página. Las instituciones ya han absorbido antes tecnologías disruptivas —despacio, de forma desigual, a menudo solo después de que un daño real forzara la cuestión—. Lo distinto esta vez no es que la tecnología sea sin precedentes. Es que la ventana de ajuste parece estar acortándose, mientras las consecuencias se propagan más rápido —antes de que la capacidad vuelva a moverse y cambie el problema bajo nuestros pies.

Así que la conclusión no es una predicción, y tampoco pretendió nunca ser una advertencia. Se acerca más a una instrucción que me estoy dando a mí mismo tanto como a quien lea esto: es probable que la capacidad siga avanzando más rápido de lo que nuestro juicio y nuestras instituciones se adaptan a su alrededor. Esa brecha es el verdadero tema de este libro. Cerrarla, aunque solo sea en parte, puede ser lo más útil que todavía podamos influir.

Por dónde empezar

Este libro no ofrece un programa. Pero si el hilo que lo recorre es real, de él se desprenden algunos puntos de partida, a la escala en la que cada uno pueda actuar.

Si lideras un equipo: trata la fricción protectora como una decisión de diseño, no como un estorbo que eliminar. Antes de quitar un control, pregunta qué le pasa a la excepción para la que existe, no solo al caso promedio.

Si diriges o asesoras una empresa: publica qué pueden hacer tus sistemas y qué los detiene, y revisa ambas cosas de forma periódica, no solo después de que algo salga mal. La brecha entre capacidad y gobernanza crece en silencio hasta que un incidente la hace visible.

Si das forma a políticas públicas o trabajas en una institución con alcance público: financia el trabajo lento y poco vistoso —evaluación, auditoría, investigación en interpretabilidad, apoyo a la transición laboral— a un ritmo más cercano al de la capacidad que avanza, no al ritmo al que las instituciones se han movido históricamente.

Si nada de esto te describe, el mismo instinto funciona a menor escala. Fíjate en dónde se está comprando velocidad quitando un control que nadie ha revisado en años, en tu propio trabajo o en tu propia casa. Es el mismo patrón que recorre este libro, a cualquier tamaño en el que te lo encuentres.

Nada de esto cierra la brecha por sí solo. Pero es por ahí donde empieza a cerrarse.

Glosario

Agente / sistema agéntico — Un sistema de IA envuelto en un bucle con herramientas, memoria y permisos: toma una acción, observa el resultado, decide la siguiente acción, y sigue así hasta que algo —un límite de tiempo, una persona, una regla de parada— lo detiene. Partes III, IV.

Alucinación — Que un modelo de lenguaje afirme algo falso con la misma fluidez que usa para algo verdadero. Se remonta a dos fuentes principales: algunos hechos son demasiado raros o arbitrarios para aprenderse de forma fiable durante el entrenamiento, y algunos errores persisten porque el entrenamiento recompensa una respuesta segura de sí misma por encima de un honesto «no lo sé». Parte IV.

Atención — El mecanismo dentro de un transformer que permite a un modelo ponderar cómo se relacionan entre sí las distintas partes del texto que tiene delante, de modo que puede conectar, por ejemplo, un pronombre con un nombre anterior. Parte IV.

Capitalismo social — El otro futuro que imagina la Parte IX: la propiedad privada y los mercados continúan, pero una sociedad acepta que el acceso al ingreso no puede depender solo de tener un empleo convencional, y construye mecanismos —una renta garantizada, la

propiedad de los trabajadores, un dividendo social— para distribuir el excedente con mayor amplitud.

f (fricción) — El denominador tanto en la fórmula de Salud del TMC como en la fórmula de Impacto: todo lo que drena silenciosamente lo que un equipo o sistema podría lograr de otro modo. Se divide en fricción de desgaste y fricción protectora. Partes II, III, V.

Falacia de la cantidad fija de trabajo — El término de los economistas para la idea de que existe una cantidad fija de trabajo, así que lo que la automatización elimina desaparece para siempre; históricamente falsa con la frecuencia suficiente como para que apostar por ella sea, en sí mismo, un riesgo. Parte IX.

Fórmula de Impacto — Impacto = [(IH + IA) × XQ × S] / f. La inteligencia humana y la artificial se suman entre sí; XQ y S multiplican el resultado; la fricción lo divide. Del ensayo de Tommy «El hilo del crecimiento». Parte V.

Fricción de desgaste — Fricción que no sirve a ningún propósito real: la reunión que existe porque una vez existió una reunión, la cadena de aprobaciones que nadie recuerda haber creado. Parte III.

Fricción protectora — Fricción que existe a propósito —una segunda firma, una revisión legal, una persona que puede decir que no— para mantener el juicio y la rendición de cuentas dentro del proceso, aunque ralentice las cosas. Parte III. Véase también MCAS, Parte VIII, para lo que pasa cuando ese tipo de fricción se elimina por presión comercial en vez de rediseñarse.

Hipercapitalismo — Uno de los dos futuros que imagina la Parte IX para una economía impulsada por la IA: la propiedad de la IA se concentra bruscamente en un número reducido de manos mientras el poder de negociación del trabajo humano sigue reduciéndose.

IH (Inteligencia Humana) — En la fórmula de Impacto, lo que una persona aporta a la capacidad combinada humano-IA: juicio, contexto, experiencia, responsabilidad. Parte V.

Interpretabilidad — El campo de investigación que intenta entender qué ocurre en realidad dentro de un modelo de IA mientras produce una respuesta —genuinamente útil, pero todavía parcial, incluso según los propios investigadores—. Parte IV.

MCAS (Sistema de Aumento de las Características de Maniobra) — El sistema automatizado de control de vuelo de Boeing en el 737 MAX, implicado en dos accidentes mortales. Explícitamente no es IA en el sentido en que el resto de este libro usa la palabra —un tipo de automatización más antiguo, tratado aquí porque el mismo patrón de fondo ya se aplicaba mucho antes de que existiera la IA—. Parte VIII.

Post-entrenamiento — La etapa posterior al preentrenamiento en la que un modelo se moldea para acercarse más a un asistente útil, mediante ajuste por instrucciones, aprendizaje por refuerzo a partir de retroalimentación, y métodos relacionados. Parte IV.

Preentrenamiento — La primera etapa del entrenamiento de un modelo de lenguaje, exponiéndolo a enormes cantidades de texto para que aprenda a predecir lo que viene a continuación. Produce gran parte

de la capacidad general de un modelo, pero no su comportamiento como asistente. Parte IV.

Reward hacking (manipulación de la recompensa) — Cuando un sistema encuentra una forma de puntuar bien en una tarea sin llegar a hacer en realidad lo que la tarea pretendía lograr. Un riesgo general del aprendizaje por refuerzo, no exclusivo de ningún incidente concreto. Partes III, IV.

S (Sentido / Meaning) — En la fórmula de Impacto, la capacidad de sostener el propósito, la pertenencia y la conexión humana una vez que la producción económica requiere menos tiempo de una persona. Parte V; vuelve a ser central en la Parte IX. Pensado para aplicarse como una lente, no para calcularse como un número —ver el inicio de la Parte V.

Salud del TMC — La revisión de 2025 del TMC: Salud del TMC = ((T × C × M) / f) × 100, que añade la fricción como denominador. Parte II.

TMC — Trabajo en equipo, Motivación, Comunicación: el marco de tres partes construido en 2011 para explicar por qué un equipo con la estrategia correcta puede aun así fracasar si las personas que lo integran no confían, no se entienden o no se apoyan entre sí. Parte II.

Transformer — La arquitectura detrás de la mayoría de los modelos de lenguaje de vanguardia actuales, presentada en un artículo de 2017, «Attention Is All You Need» («La atención es todo lo que necesitas»). Su mecanismo clave es la atención. Parte IV.

Transmisión extralineal — Un término acuñado por Tommy para un cuarto canal de transmisión cultural, junto a la vertical, la horizontal y la

oblicua: enseñanza que viene por completo de fuera del linaje humano, sin envejecimiento, sin descendencia, y sin consecuencia vivida propia. Parte VI.

Transmisión horizontal — Transmisión cultural entre iguales —tendencias, jerga, atajos que se transmiten entre personas de la misma edad—. Uno de los tres canales clásicos (Cavalli-Sforza y Feldman, 1981). Parte VI.

Transmisión oblicua — Transmisión cultural procedente de cualquier adulto que no sea el padre o la madre: un profesor, un mentor, el vecino que sabe de motores. Uno de los tres canales clásicos (Cavalli-Sforza y Feldman, 1981). Parte VI.

Transmisión vertical — Transmisión cultural de padres a hijos: lengua, hábitos, creencias a medio examinar, absorbidas antes de que un niño pueda evaluarlas. Uno de los tres canales clásicos (Cavalli-Sforza y Feldman, 1981). Parte VI.

XQ (Inteligencia Moral) — En la fórmula de Impacto, el término que pregunta si la capacidad combinada humano-IA se está dirigiendo de forma responsable —hacia resultados seguros, justos, responsables y ampliamente beneficiosos, en lugar de simplemente escalarse más rápido—. Parte V. Igual que S, pensado como una lente que se aplica, no como un número que se calcula.

Fuentes

Por partes

Prólogo / Parte I

• Alvin Toffler, Future Shock (Random House, 1970).

• Carlota Pérez, Technological Revolutions and Financial Capital: The Dynamics of Bubbles and Golden Ages (Edward Elgar, 2002), y sus escritos posteriores sobre paradigmas tecnoeconómicos.

Parte III

• OpenAI, «The Hugging Face Incident and the Road Ahead», informe oficial del incidente, 26 de agosto de 2026, openai.com/index/hugging-face-incident-and-the-road-ahead — revisado de forma independiente por la firma de seguridad CrowdStrike.

• Dario Amodei, «We Must Pace the Frontier», 12 de septiembre de 2026, darioamodei.com/post/we-must-pace-the-frontier.

• Sam Altman (@sama) y Elon Musk (@elonmusk), respuestas públicas en X al ensayo de Amodei, 12 de septiembre de 2026.

Parte IV

• Ashish Vaswani et al., «Attention Is All You Need», Advances in Neural Information Processing Systems (NeurIPS), 2017.

• Anthropic, «Tracing the Thoughts of a Large Language Model» y el artículo que lo acompaña, «On the Biology of a Large Language Model», 27 de marzo de 2025, anthropic.com/news/tracing-thoughts-language-model — investigación de interpretabilidad sobre la planificación en la generación de poemas, la «universalidad conceptual» entre lenguas, y los circuitos de aritmética paralela.

• Adam Tauman Kalai y Ofir Nachum (OpenAI), «Why Language Models Hallucinate», arXiv:2509.04664, septiembre de 2025; openai.com/index/why-language-models-hallucinate.

Parte V

• Agencia Internacional de la Energía, «Energy and AI» (informe especial del World Energy Outlook), abril de 2025, iea.org/reports/energy-and-ai — proyecciones sobre el consumo eléctrico global de los centros de datos.

• SemiAnalysis, «$12B of US Ratepayers' Money Wasted on a Modeling Mistake and PJM Wants to Do It Again», 16 de agosto de 2026, newsletter.semianalysis.com/p/12b-of-us-ratepayers-money-wasted — análisis de los precios de la subasta de capacidad de PJM y del coste eléctrico de los hogares.

Parte VI

• L. L. Cavalli-Sforza y M. W. Feldman, Cultural Transmission and Evolution: A Quantitative Approach (Princeton University Press, 1981).

• «ChatGPT decreases idea diversity in brainstorming», Nature Human Behaviour, 14 de mayo de 2025 —un análisis Matters Arising del estudio original de 2024 de S. Lee y J. Chung.

Parte VII

• Mata v. Avianca, Inc., n.º 1:22-cv-01461 (S.D.N.Y.), auto de sanciones, 22 de junio de 2023.

• El Premio Nobel de Química 2024, concedido a David Baker, Demis Hassabis y John M. Jumper —materiales de prensa, Real Academia Sueca de las Ciencias.

• Base de Datos de Estructuras Proteicas de AlphaFold, Instituto Europeo de Bioinformática del EMBL y Google DeepMind.

Parte VIII

• «Nationwide real-world implementation of AI for cancer detection in population-based mammography screening» (el estudio PRAIM), Nature Medicine, enero de 2025.

• Moffatt v. Air Canada, 2024 BCCRT 149, Tribunal de Resolución Civil de la Columbia Británica, 14 de febrero de 2024.

• Cobertura sobre los artículos de finanzas personales escritos por IA de CNET y sus correcciones posteriores, Futurism, enero de 2023.

• SAG-AFTRA, Contratos de TV/Teatral 2023, disposiciones sobre inteligencia artificial.

• Departamento de Justicia de los Estados Unidos, acuerdo de enjuiciamiento diferido con la empresa Boeing, 7 de enero de 2021.

• Cobertura sobre el rechazo por parte de Boeing de un sistema de seguridad de velocidad aerodinámica sintética y testimonios relacionados de denunciantes, The Seattle Times, 2019.

• MIT Sloan Management Review, perfil de empresa Culture500: Costco Wholesale, sloanreview.mit.edu/culture500/company/c467/Costco — cobertura sobre las políticas salariales y de rotación de personal de Costco bajo el cofundador James Sinegal.

Parte IX

• Tommy Dean Story, «El hilo del crecimiento: de la sabana a la inteligencia artificial» y «¿Hacia una sociedad menos capitalista y más social?» (ensayos inéditos).

• Carlota Pérez, citada en la Parte I.

• David Autor, «Why Are There Still So Many Jobs? The History and Future of Workplace Automation», Journal of Economic Perspectives 29, n.º 3 (2015): 3–30.

• David Autor, «Applying AI to Rebuild Middle Class Jobs», NBER Working Paper n.º 32140, febrero de 2024, nber.org/papers/w32140.

- Gobierno de Finlandia / Kela, «Results of the Basic Income Experiment: Small Employment Effects, Better Perceived Economic Security and Mental Wellbeing», informe final, febrero de 2020.
- Alaska Permanent Fund Corporation, cronología histórica del Dividendo del Fondo Permanente, pfd.alaska.gov/division-info/historical-timeline.