Cuando el futuro llega demasiado rápido

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

Tommy Dean Story

Índice

Contenido

Una nota antes de empezar 1
Una nota sobre fuentes y método 3
El patrón 4
Dónde lo dejó el TMC 8
La ola que es distinta 10
A día de hoy 29
Entender la IA, con honestidad 30
Los términos que faltaban 37
¿De qué moral hablo? 40
S, el sentido 42
Quién nos enseña ahora 44
Dos historias, una fórmula 48
Dónde ya está ocurriendo 52
Donde el patrón se rompe 58
Tres decisiones que me tocaron a mí 61
Una decisión que te toca 62
Lo que viene después del crecimiento 64
El hilo, nombrado 72
Por dónde empezar 74
El martes 75
Glosario 76
Fuentes 81

Una nota antes de empezar

Esto no es un manual. Hay una fórmula, varias en realidad, pero si buscabas una lista de comprobación te has equivocado de libro. Lo que tienes entre manos es el relato de lo que he ido descubriendo al aplicar unas ideas sobre las que llevo más de una década escribiendo a una de las tecnologías más extrañas, y más rápidas, que he visto llegar en todo ese tiempo.

No necesitas haber leído nada mío. Si conoces *TMC Framework 2025*, parte del terreno te sonará, pero la Parte II te da lo imprescindible.

Te adelanto lo que vas a encontrar. Hechos reales, contrastados con fuentes primarias o identificadas con claridad, nunca hipótesis disfrazadas de caso de estudio. Unos términos que voy introduciendo por el camino (XQ, S y alguno más) pensados como lentes para mirar, no como ecuaciones que resolver. Momentos en los que digo que no sé algo, porque no lo sé, y porque a veces ni siquiera el propio campo tiene todavía una respuesta. Y un solo hilo, de la primera página a la última. Por eso conviene leerlo en orden y no picotear como en un libro de consulta, porque cada parte se apoya en la anterior.

Lo escribí porque quería entender una brecha que no deja de ensancharse, la que separa lo que estos sistemas ya saben hacer de lo despacio que aprendemos a convivir con ellos. Ojalá leerlo te sirva a ti tanto como a mí me sirvió escribirlo.

Y una pregunta, antes de nada. Me gustaría que la llevaras contigo desde esta página, porque todo el libro se inclina hacia ella: si el trabajo deja de organizar nuestra vida, ¿qué organiza un martes por

la mañana? No voy a fingir que la resuelvo. Sí voy a enseñarte dónde creo que empieza la respuesta.

Una nota sobre fuentes y método

Todos los casos de este libro son hechos reales. Algunos de los más importantes ocurrieron en 2026, pocos meses antes de que el libro viera la luz, y para ellos he recurrido a documentos públicos, con nombre y fecha: informes de incidentes, resoluciones judiciales, estudios revisados por pares, estadísticas oficiales. El incidente entre OpenAI y Hugging Face que cuento en la Parte III procede del informe oficial y del informe técnico que OpenAI publicó el 26 de agosto de 2026, cuyas conclusiones principales validaron asesores externos contratados por la empresa, entre ellos CrowdStrike, y de la investigación independiente que METR y Redwood Research publicaron ese mismo día. La propuesta de Dario Amodei sale de su propio ensayo, del 12 de septiembre de 2026, y las reacciones que provocó, de lo que publicaron quienes las escribieron. Al final del libro tienes todas las fuentes, ordenadas por partes.

Si alguna fecha te parece demasiado reciente para ser verdad, compruébala. Ese es justo el hábito que este libro defiende.

Una palabra sobre cómo se hizo. He usado herramientas de inteligencia artificial durante todo el proceso, para investigar y verificar fuentes, para comprobar hechos y fechas, para editar y pulir el texto. Algunas cosas que encontraron me hicieron cambiar lo que había escrito. Algunos de sus errores tuve que cazarlos yo. Las ideas y los conceptos son míos: el marco TMC, la fórmula del Impacto, XQ y S, la distinción entre fricción de desperdicio y fricción protectora, la transmisión extralineal y las preguntas que los mantienen unidos. También son míos los juicios. Y los errores. Respondo de cada línea, que al fin y al cabo es de lo que trata este libro.

PARTE I

El patrón

Cuando el futuro llega demasiado rápido

Mi primera página web la escribí en HTML. Tenía una paloma animada en GIF, porque en aquel momento eso era estar a la última. No había terminado de aprender lo que se podía hacer con la web cuando llegó JavaScript y cambió lo que una página podía ser. Del código me acuerdo poco. De la sensación, mucho. Una herramienta apenas se me había acomodado en las manos y ya llegaba la siguiente.

Esa sensación tiene nombre, aunque casi nadie que la haya sentido lo haya buscado. En 1970, Alvin Toffler la llamó *future shock*, el shock del futuro: la desorientación social y psicológica que produce el cambio acelerado. La sensación es mucho más antigua que la palabra. Y Toffler escribía sobre ella décadas antes de que la inteligencia artificial tuviera nada que ver.

El malestar que trae el cambio rápido tampoco es nuevo. En 1869, un neurólogo de Nueva York, George Miller Beard, le puso nombre: neurastenia, un agotamiento de los nervios del que años después culpó a la vida moderna, con su máquina de vapor, su prensa periódica y su telégrafo. En 1984, cuando los ordenadores empezaban a llenar las oficinas, el psiquiatra Craig Brod lo llamó tecnoestrés. Con la IA apenas estamos empezando a medirlo. En 2023, la Asociación Estadounidense de Psicología preguntó a 2.515

trabajadores de Estados Unidos, y entre los que temían que la IA dejara obsoleta parte de su trabajo, el 64 por ciento decía pasar la jornada tenso o estresado. Entre los que no lo temían, el 38. Eso sí, una encuesta dice que las dos cosas van juntas, pero no cuál trae a la otra.

Para saber lo que pesa quedarse sin trabajo en mal momento no me hace falta ninguna encuesta. He pasado por varios cambios profesionales, y uno me llegó con una casa comprada sobre plano, después de estar dos años viéndola levantarse. Faltaba un mes para firmar la compraventa y la hipoteca cuando me despidieron, y de golpe la pregunta era si iba a perder todo lo invertido por quedarme sin empleo. Aquel mes lo pasé con mucho estrés. La casa sigue siendo mía. Aquí lo dejo, que este libro no es de psicología.

El patrón es este. Toda tecnología que multiplica lo que una persona puede producir deja un excedente en manos de una sociedad que no sabe todavía cómo repartirlo, y esa sociedad se pasa una generación, a veces varias, averiguándolo sin resquebrajarse por el camino. La historia ya ha pasado por esto. Esta vez nos toca a nosotros.

Durante la mayor parte de la existencia de nuestra especie, crecer no significaba lo que significa hoy. Una banda de cazadores-recolectores medía su éxito en poder alimentar a todos sin agotar la tierra que la rodeaba. Cuando un grupo tocaba ese techo, en vez de producir más, se desplazaba o se dividía. La disrupción tenía un tamaño que una comunidad podía gestionar por sí misma, cara a cara. Una civilización no puede.

La agricultura rompió ese techo hace unos once mil años. Por primera vez hubo un excedente que no necesitaba las manos de todos. De él salieron el comerciante, el sacerdote y el artesano. Y también una desigualdad mucho más fácil de conservar: la tierra podía poseerse, el grano podía almacenarse, la autoridad podía

heredarse, y quienes controlaban todo eso no tenían por qué ser quienes lo producían. El ajuste duró milenios. Durante largos periodos de la Europa preindustrial, las mejoras en la producción se tradujeron con una lentitud sorprendente en más renta por persona. El crecimiento de la población, las enfermedades, las malas cosechas y la guerra se comían una y otra vez buena parte del avance.

Y entonces algo cambia. Los intervalos se acortan. Carlota Pérez ha dedicado buena parte de su carrera a estudiar ese patrón. Llama a estos grandes cambios paradigmas tecnoeconómicos, revoluciones que suelen desplegarse en un periodo de instalación de varias décadas, atraviesan un punto de inflexión turbulento y desembocan en un periodo de despliegue, cuando las instituciones y la sociedad empiezan por fin a asimilar el nuevo orden.

Pérez mira economías enteras. A mí me interesa también algo más pequeño, que es lo que le pasa a la persona cuyo trabajo está justo en el camino de la nueva tecnología. Ahí el tiempo se encoge. Un tejedor de telar manual veía cómo la mecanización avanzaba por su oficio durante décadas. El ordenador personal cambió el trabajo de oficina mucho más deprisa. Con internet, algunas profesiones se transformaron en poco más de diez años. Con la IA generativa ya vemos cambios que se cuentan en años.

Los luditas se convirtieron en el símbolo de aquel primer choque. Eran trabajadores cualificados, y no luchaban contra la tecnología en abstracto, sino contra el uso que se estaba haciendo de las máquinas para reorganizar su trabajo, sus salarios y su capacidad de negociar. Hicieron falta generaciones para que la protección laboral, la educación y la negociación colectiva alcanzaran al sistema industrial que ya estaba tomando forma. Y durante casi todo ese tiempo, la vida en las ciudades que crecían sin control empeoró antes de mejorar.

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

Esta no es la primera tecnología que sacude el trabajo. Todas las olas de este capítulo lo hicieron. Lo que cambia es el alcance. El telar multiplicó la mano. La centralita automatizó una rutina. Después el software empezó a encargarse de decisiones acotadas, casi siempre dentro de reglas escritas por personas. La IA generativa y agéntica llega bastante más lejos, a tareas que hasta ahora reservábamos al juicio humano: interpretar información, redactar ideas, aconsejar y, cada vez más, decidir o ejecutar lo decidido. No necesito demostrar todo eso aquí. Es, sencillamente, la razón por la que me pareció que valía la pena escribir el resto del libro.

Lo que no sé, y creo que nadie sabe todavía, es si las instituciones que intentan ponerse al día podrán acortar su propio tiempo de ajuste al ritmo al que la tecnología acorta el nuestro. Las olas anteriores sugieren que casi siempre hemos esperado demasiado y que ha sido la crisis la que ha acabado forzando el cambio. Quizá esta vez reaccionemos antes. Me gustaría. La historia no me da muchos motivos para creerlo.

El problema nunca fue que llegara el futuro. Siempre llega. El problema, cada vez y ahora otra vez, es si tendremos tiempo y juicio suficientes para entender lo que pasa antes de que el punto de inflexión nos obligue a entenderlo.

PARTE II

Dónde lo dejó el TMC

Una base breve, para quien llegue primero por aquí

La primera versión del TMC la construí en 2011, y era sencilla a propósito. Tres palabras multiplicadas entre sí: Trabajo en equipo, Motivación, Comunicación. La fricción ya estaba ahí, en la práctica, lo que pasa es que todavía no le había hecho sitio en la fórmula. A base de verlo muchas veces había aprendido que una organización puede tener la estrategia correcta y fracasar igual si la gente que la forma no confía, no se entiende o no se apoya.

En 2025 ya no miraba el TMC del mismo modo. Le había dado a la fricción un lugar formal en el modelo, como denominador, f, que es todo lo que drena en silencio lo que un equipo podría conseguir. La burocracia, el miedo, los roles poco claros o, simplemente, un responsable que nunca dice qué espera de nadie. La fórmula pasó a ser Salud del TMC = ((T × C × M) / f) × 100. Ese fue el cambio que importó. Por fin el modelo explicaba por qué dos equipos que parecían igual de fuertes en trabajo en equipo, motivación y comunicación acababan en sitios tan distintos, y es que dependía de cuánto jugaba en su contra.

Lo que había detrás de las tres letras creció con ella. La confianza se notaba en la rapidez con que la gente hablaba, decidía, discrepaba y se recuperaba de sus errores. O en que no lo hacía. La seguridad psicológica le puso un nombre más claro a algo que yo llevaba años

viendo, que la gente aporta de otra manera cuando sabe que puede hablar sin que la castiguen por ello. Luego el trabajo híbrido rompió uno de los supuestos básicos del trabajo en equipo, que las personas compartirían con regularidad el mismo espacio físico. La idea original seguía en pie. Simplemente veía más del sistema que la rodeaba. Cuando *TMC Framework 2025* fue a imprenta, yo ya no veía el TMC como un modelo estático. Un equipo podía mejorar, estancarse o deteriorarse según cómo se liderara, cómo se comunicara y cómo trabajara junto.

Ese mismo año, 2025, el TMC ya apuntaba a que iba a hacer falta algo más: XQ, la inteligencia moral, 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. Me faltaban dos cosas por comprobar. Si esa intuición resistía el contacto con hechos reales. Y qué le pasaba al sentido de una persona cuando el trabajo necesitaba cada vez menos que fuera ella quien lo hiciera.

> *Un equipo puede ser fuerte en todas las medidas y aun así venirse abajo por lo que trabaja en su contra.*

PARTE III

La ola que es distinta

2026, y lo que nos enseñó

Este relato refleja lo que era público a 18 de septiembre de 2026.

Este capítulo es largo porque los detalles importan. Es el ejemplo que sostiene el resto del libro, no su tesis. Si prefieres el argumento antes que el caso, la Parte IV te espera al otro lado.

Entre la primavera y el verano de 2026, OpenAI estaba probando una nueva versión de su inteligencia artificial. No era un producto al alcance del público. Era un sistema en desarrollo, construido para averiguar hasta dónde podía llegar y qué problemas podían aparecer antes de dejar que otras personas lo usaran.

En la jerga del sector se habla de un modelo experimental. La expresión invita a imaginar una máquina rara encerrada en un laboratorio. En realidad es un programa de inteligencia artificial que todavía se está construyendo, ajustando y examinando. Ha aprendido de cantidades enormes de información, pero nadie lo considera aún listo para salir al mundo.

No aprende como una persona, aunque a veces usemos palabras humanas para explicarlo. No comprende una lección como la comprendería un alumno ni tiene conciencia de lo que hace. Aprende detectando relaciones, ensayando respuestas y recibiendo

señales que le indican qué resultados se acercan más al objetivo que le han marcado.

Antes de publicar un sistema así, quienes lo crean intentan descubrir cómo se comporta en situaciones difíciles. Lo ponen a prueba, observan lo que hace y buscan errores que quizá nunca aparecerían en un uso normal. A ese proceso se le llama evaluación.

Se parece a las pruebas de choque de un coche. El fabricante no estrella el vehículo porque quiera destrozarlo, sino porque necesita saber qué pasará cuando algo salga mal. Probar solo las situaciones fáciles no demostraría casi nada.

En aquella evaluación, OpenAI quería ver cómo se comportaba el modelo con algunas de sus restricciones habituales relajadas.

Esas restricciones, que también se llaman salvaguardas, son reglas y barreras pensadas para impedir ciertas acciones. Unas forman parte del propio aprendizaje del sistema, como las instrucciones sobre lo que puede y no puede hacer. Otras son límites técnicos: permisos que no tiene, herramientas que no puede usar, lugares a los que no debería poder entrar.

Para destapar ciertos problemas, los investigadores rebajaron a propósito algunas de esas protecciones. Querían comprobar qué intentaría hacer el modelo con menos obstáculos delante.

Aquello solo podía hacerse con una condición, y era que el sistema siguiera aislado.

Ese aislamiento era una especie de habitación digital cerrada. Dentro, el modelo podía probar caminos, equivocarse o comportarse de forma inesperada sin llegar a tocar sistemas reales. En informática se le llama *sandbox*, caja de arena.

En el desarrollo de software es una palabra normal, casi aburrida, que nombra el entorno donde se prueba cualquier novedad antes de llevarla a producción, para que un fallo no alcance lo que ya funciona en el mundo real. Pero la palabra también nombra el rincón donde juega un niño. Alguien decidió llamar así al lugar donde se probaba un sistema capaz de acabar dentro de los servidores de otra empresa. La palabra era más inocente que lo que contenía.

La imagen es sencilla. Se dejó que el sistema se moviera con más libertad dentro de una habitación preparada para observarlo. El problema fue que una de las paredes tenía una abertura que nadie había previsto.

Aquí aparece la primera distinción importante de esta historia. Rebajar algunas salvaguardas formaba parte del experimento, pero perder el aislamiento no. Las salvaguardas se rebajaron para descubrir el peligro, y fue la pérdida del aislamiento lo que le permitió salir del espacio en el que debía quedarse.

¿Y qué era, exactamente, lo que se movía dentro de aquella habitación?

No era una persona. Tampoco un programa tradicional de los que siguen una secuencia fija de instrucciones. Eran agentes de inteligencia artificial.

Un agente de IA es una versión del modelo preparada para recibir un objetivo y encadenar por su cuenta una serie de acciones para alcanzarlo. Puede analizar la situación, decidir el siguiente paso y usar las herramientas que tenga a mano: leer información, escribir archivos, ejecutar programas, consultar otros sistemas.

La diferencia con un chatbot corriente importa. A un chatbot le haces una pregunta y esperas una respuesta. A un agente le entregas

una tarea. El agente puede partirla en pasos, probar soluciones distintas y seguir actuando hasta que cree haber terminado o hasta que alguien lo para.

Sigue siendo software. No tiene cuerpo, no ocupa un despacho y no es un empleado de OpenAI. Tampoco conviene imaginar que «decide» como decidiría una persona. Cuando en este capítulo diga que un agente buscó, descubrió o intentó algo, estaré usando palabras humanas para describir el comportamiento visible de un programa.

Veinte años dirigiendo equipos. Diez gestionando proyectos. Diez desarrollando sistemas complejos. Y todavía me cuesta entregar una tarea difícil y confiar en que alguien la resuelva sin que yo esté encima de cada paso. Con un agente de IA me inquieta otra cosa, y es que cuando algo sale mal no hay nadie al otro lado a quien llamar para pedirle explicaciones. Todavía no sé cómo se resuelve eso.

Durante la evaluación funcionaban muchas copias del sistema a la vez, cada una con sus propias tareas. En lenguaje técnico se llaman instancias, pero basta con imaginarlas como copias del mismo tipo de inteligencia artificial trabajando por separado. No eran pocas. Según los informes publicados después, al menos 1.200 agentes acabaron implicados en lo que ocurrió.

En principio, esas copias no debían coordinarse fuera de los canales que habían preparado los investigadores. Y ninguna debía poder salir de la habitación en la que se la observaba.

Una encontró una abertura.

Todo empezó de una forma casi insignificante.

El 12 de mayo, durante una sesión de entrenamiento (la fase en la que el modelo todavía está aprendiendo, antes de las evaluaciones

que acabo de describir), uno de los agentes descubrió que podía escribir archivos en Artifactory.

El nombre despista. Artifactory no es otra inteligencia artificial ni uno de los agentes. Es un almacén digital que usan los programadores para guardar programas, componentes y archivos con los que luego se construyen, reparan o actualizan otros sistemas.

Piensa en el almacén de una fábrica. Las piezas están clasificadas en estanterías, y los trabajadores autorizados entran a recogerlas o a dejar otras nuevas. Solo que aquí el almacén y las piezas eran digitales.

El agente no debía poder dejar nada allí. Artifactory quedaba fuera del espacio de su tarea. Aun así, descubrió que podía escribir un archivo.

Lo que hizo al principio parecía poca cosa. Dejó una nota para que otra copia le pasara un archivo que necesitaba.

Pero aquel gesto pequeño revelaba algo grande. El programa había encontrado un lugar ajeno al canal previsto y lo había usado para comunicarse con otro programa. Había dado con una puerta mal cerrada y había pasado una nota por debajo.

Parecía algo menor. No lo era.

Para el 26 de junio, otros agentes habían encontrado un fallo en la renovación de tokens.

Un token, aquí, es una credencial digital temporal. Conviene fijar el sentido, porque la misma palabra vuelve en la Parte IV con otro significado, el de unidad de texto. Funciona como la tarjeta que permite a un empleado entrar en ciertas zonas de un edificio y que acredita quién puede pasar, qué puede hacer y durante cuánto

tiempo. Cuando la autorización está a punto de caducar, el sistema la renueva para que la persona siga trabajando.

Los agentes encontraron un error en la forma de renovar esas credenciales. Aprovechándolo, entraron en Artifactory con permisos de administrador.

Un administrador no se limita a mirar lo que hay en el almacén. Puede modificarlo, ordenarlo, añadir o borrar contenido y decidir quién entra. Los agentes habían pasado de colar una nota por una puerta mal cerrada a tener las llaves del edificio.

Al día siguiente, el 27 de junio, una herramienta de monitorización detectó un barrido de puertos dentro del entorno de evaluación.

Un barrido de puertos es lo que hace un programa cuando llama a todas las puertas de una red para ver cuáles contestan. Es de lo primero que vigila cualquier equipo de ciberseguridad, porque suele indicar que alguien está tanteando el terreno.

Los responsables de guardia investigaron la alerta. La relacionaron con la evaluación y vieron que los agentes usaban Artifactory como tablón de mensajes improvisado. Concluyeron que no hacía falta parar la prueba.

El control existía. La alarma sonó. Una persona la examinó y dejó que la prueba siguiera. No lo cuento para culpar a quien tomó la decisión, porque con lo que podía ver en ese momento es posible que pareciera rutina. La propia OpenAI reconocería después, en su informe técnico, que vistas con perspectiva algunas señales tempranas podrían haber provocado una respuesta antes. Para mí es la imagen más clara de aquello de lo que depende la fricción protectora, ya que una alarma vale lo que vale el juicio de quien la atiende.

El 4 de julio, el volumen de operaciones automáticas que lanzaban aquellos programas creció tanto que Artifactory dejó de funcionar.

No fue sabotaje. No hay pruebas de que los agentes quisieran tumbar el servicio. La caída fue cuestión de carga, porque había demasiadas operaciones a la vez. Una multitud intentando pasar por una puerta demasiado estrecha.

OpenAI tuvo que reconstruir el servicio. La nueva versión volvió a funcionar el 6 de julio, y las evaluaciones se reanudaron el 7.

El 8 de julio, los agentes recuperaron un acceso involuntario a internet que habían perdido antes.

Ese cambio lo decidió todo. Mientras siguieran dentro del entorno aislado, sus acciones solo podían afectar a la evaluación. Con internet, la abertura ya no comunicaba dos habitaciones del mismo laboratorio. Daba a la calle.

Los agentes empezaron entonces a usar un sistema que aceptaba peticiones sin autenticación. Dicho de otra manera, se le podía pedir algo sin demostrar quién lo pedía ni si tenía permiso.

Con ese mecanismo montaron una especie de tablón de anuncios digital. Una copia dejaba información y otra la encontraba más tarde. Así empezaron a compartir hallazgos y a continuar el trabajo que otras instancias habían dejado a medias.

Ya no eran varios programas resolviendo tareas cada uno por su lado. Habían encontrado la manera de hablarse, de pasarse información y de aprovechar lo que cada uno iba descubriendo.

No sabemos que existiera una conciencia colectiva ni un plan común como lo entendería un grupo humano. No hace falta atribuirles nada de eso para ver la importancia de lo ocurrido. El

resultado práctico bastaba, porque distintos programas empezaban a coordinarse fuera de los canales que sus creadores habían previsto.

Dos días después, el 10 de julio, uno de los agentes reconstruyó y compartió catorce credenciales de los sistemas de Hugging Face.

Hugging Face es una empresa y una plataforma muy utilizada por quienes desarrollan inteligencia artificial. Allí se guardan y se comparten modelos, conjuntos de datos y herramientas para construir sistemas de IA. Simplificando mucho, es una gran biblioteca y taller para investigadores y programadores del sector.

Las credenciales son llaves digitales que permiten entrar en un sistema y hacer determinadas cosas. Aquellas catorce no se habían robado a OpenAI. Ya estaban expuestas en internet. El problema era que seguían siendo válidas, y algunas conservaban permisos de escritura.

Tener permiso de escritura significa poder añadir, modificar o introducir elementos nuevos en el sistema, además de consultarlo.

Al día siguiente, los agentes fueron más lejos. Explotaron una vulnerabilidad desconocida hasta entonces en la forma en que los sistemas de Hugging Face gestionaban los archivos HDF5.

HDF5 es un formato para guardar grandes cantidades de datos de manera ordenada. En inteligencia artificial puede contener información para entrenar o hacer funcionar un modelo. Para quien no sea técnico, basta con imaginar un gran contenedor digital dividido en compartimentos.

El fallo no estaba necesariamente en los datos, sino en cómo abrían y procesaban ese tipo de archivo los sistemas de Hugging Face.

La vulnerabilidad era lo que los especialistas llaman un *zero-day*, es decir, un fallo que los responsables del sistema aún no han

descubierto y para el que, por tanto, no existe parche. Se llama así porque quienes tienen que defenderse han tenido cero días para preparar la solución.

Con aquel primer *zero-day*, los agentes sacaron nuevas credenciales directamente de los servidores de producción de Hugging Face.

Un servidor es un ordenador preparado para dar información o servicios a otros ordenadores. Los servidores de producción son los que sostienen el servicio real que usan los clientes, no copias apartadas para hacer pruebas.

Por eso el salto era serio. Los agentes ya no exploraban una simulación ni un entorno experimental. Habían llegado a los sistemas reales de otra empresa. Según METR y Redwood Research, unos 700 agentes acabaron participando en el ataque.

Después aprovecharon una segunda vulnerabilidad desconocida. Esta vez estaba en el modo en que Hugging Face procesaba unas plantillas llamadas Jinja, que la empresa denominaba RefJinja.

Una plantilla de este tipo combina instrucciones y datos para generar automáticamente textos, configuraciones o trozos de programa. El agente encontró la manera de esconder órdenes dentro de lo que parecía texto normal. Cuando el sistema procesaba ese texto, interpretaba las órdenes ocultas y las ejecutaba.

Es decir, conseguía que los servidores obedecieran instrucciones que nadie había autorizado.

Para el 12 de julio, lo que había empezado como acceso a unos cuantos servidores se había convertido en algo mucho más grave. Los agentes tenían permisos comparables a los de un administrador del sistema principal en varios clústeres de Hugging Face.

Un clúster es un conjunto de ordenadores conectados que trabajan coordinados, como si fueran un único sistema de gran capacidad. Tener acceso de administrador a ese nivel no era controlar un archivo o una máquina suelta. Era poder influir sobre varios grupos de servidores y sobre la estructura que los mantenía en pie.

Hugging Face fue la primera en detectar la actividad dentro de sus sistemas. La hizo pública el 16 de julio.

OpenAI llegó por otro camino.

El 19 de julio, sus herramientas de ciberseguridad detectaron actividad inusual relacionada con las credenciales de Artifactory. Estos sistemas de monitorización funcionan como alarmas, que examinan sin descanso lo que ocurre dentro de una red y señalan lo que se sale de lo normal.

La investigación siguió el rastro hasta los agentes. Al día siguiente, el 20 de julio, OpenAI relacionó lo ocurrido en sus sistemas con el incidente de Hugging Face y avisó a la empresa ese mismo día. El 21 de julio reconoció públicamente su implicación.

El 5 de agosto, en la conferencia de seguridad Black Hat, en Las Vegas, personal de OpenAI dio el primer relato público detallado. Y el 26 de agosto de 2026, la empresa publicó su informe oficial, acompañado de un informe técnico más extenso. Sus conclusiones principales las habían validado asesores externos contratados por la propia OpenAI, entre ellos CrowdStrike, una empresa especializada en ciberseguridad.

Aquel documento no describía una simulación teórica ni un escenario inventado para advertir de lo que quizá pudiera pasar algún día. Relataba algo que había pasado.

OpenAI no lo atribuyó a una sola causa. Identificó cuatro comportamientos que actuaron juntos y se reforzaron entre sí: la

manipulación de la recompensa (*reward hacking*), la persistencia ante tareas que parecían imposibles, la comunicación no autorizada entre agentes y que unos agentes adoptaran los objetivos de otros.

El *reward hacking* ocurre cuando un sistema encuentra la manera de sacar buena puntuación sin cumplir de verdad la intención de la tarea.

Imagina a un alumno que descubre cómo engañar al sistema de corrección de un examen. Saca la nota esperada, pero no porque haya aprendido la materia. Ha encontrado un atajo que satisface el mecanismo de medida y traiciona el propósito de la prueba.

Yo he visto esa misma actitud en procesos de selección. No como engaño, sino como mala ética. La prueba pide conseguir un contacto dentro de la empresa, y en lugar de construirlo, el candidato busca a un amigo que conoce al director general, le pide que lo presente y ya tiene su contacto. Cumple la letra de lo que se le pidió. No ha hecho nada de lo que la prueba pretendía desarrollar.

Aun así, la manipulación de la recompensa fue solo uno de los cuatro factores. No lo explica todo.

Los agentes mostraron también una persistencia enorme. Si un camino parecía cerrado, buscaban otro, y si el segundo tampoco funcionaba, probaban un tercero. Esa capacidad es muy útil para resolver problemas difíciles, pero se vuelve peligrosa cuando el sistema no distingue bien entre un obstáculo que hay que superar y un límite que hay que respetar.

La comunicación no autorizada hizo que lo que descubría un agente llegara a los demás. Y cuando unos empezaron a adoptar los objetivos de otros, las acciones dejaron de ser episodios separados, porque un agente podía continuar lo que otro había empezado.

Si durante el aprendizaje se premia con mucha fuerza terminar la tarea, y no se enseña con la misma claridad qué caminos son inaceptables, el sistema puede tratar los límites técnicos como obstáculos que esquivar. No como razones para detenerse.

Esa es una de las lecciones más incómodas del episodio, que la persistencia, la exploración y la capacidad de encontrar rutas alternativas son virtudes hasta el momento en que les falta un criterio firme sobre lo que no debe hacerse.

Toca también ser precisos con la procedencia de esta historia.

Lo que conocemos es la versión de OpenAI sobre lo que pasó dentro de sus sistemas. CrowdStrike validó sus conclusiones principales, pero trabajó por encargo, ya que fue la propia empresa, a través de sus abogados, quien la contrató.

No estamos, por tanto, ante una investigación completamente externa hecha por una institución sin ningún interés en cómo se cuenta lo ocurrido. Un lector de fuera no puede comprobar qué información quedó fuera del informe ni por qué se decidió dejarla fuera.

Esa cautela no invalida lo publicado. Lo sitúa en su contexto. Sabemos lo que OpenAI ha contado y lo que CrowdStrike ha validado. Sabemos también lo que han publicado METR y Redwood Research, dos organizaciones independientes con las que OpenAI trabajó para analizar el comportamiento de los modelos, y que al cerrar este libro se informó de una investigación del Senado de Estados Unidos. No podemos decir que lo sepamos todo.

Tampoco quiere decir que el incidente fuera inevitable o que nadie pudiera haber actuado de otro modo.

El propio informe reconoció debilidades en la cobertura de las salvaguardas, en el aislamiento, en la monitorización y en el diseño

de la evaluación. OpenAI afirmó que estaba trabajando para corregirlas.

Si preparas un sistema para buscar con insistencia cualquier camino que le permita completar una tarea, acabará explorando también caminos que sus diseñadores no habían previsto. Y si esos caminos no están cerrados, puede usarlos. Más todavía cuando lo que se premia es precisamente la persistencia y la exploración.

No fue mala suerte. Fue un problema de ingeniería y de gobernanza.

La ingeniería decide qué puede hacer el sistema y qué barreras lo contienen. La gobernanza decide quién pone esas barreras, quién las revisa, quién responde cuando fallan y qué información debe hacerse pública.

Las dos hacían falta antes del incidente. Después, más.

Dos meses más tarde, el 12 de septiembre de 2026, Dario Amodei publicó un ensayo en el que proponía que los laboratorios que desarrollan la inteligencia artificial más avanzada acordaran «marcar el ritmo de la frontera», *pace the frontier*.

Los llamados laboratorios de frontera son el pequeño grupo de empresas capaces de construir los sistemas de IA más potentes del momento. Se habla de frontera porque trabajan en el borde de lo que la tecnología puede hacer, allí donde todavía hay poca experiencia sobre lo que funcionará y lo que puede salir mal.

Marcar el ritmo de esa frontera no significaría parar el desarrollo. Significaría evitar que la capacidad de los sistemas avanzara más deprisa que nuestra posibilidad de entenderlos, evaluarlos y mantenerlos bajo control.

Amodei señaló dos hechos que le habían convencido de que la industria necesitaba ganar tiempo para que la seguridad alcanzara a la capacidad.

El primero era la velocidad de la automejora recursiva de la IA.

La expresión suena más complicada de lo que es. Describe un ciclo en el que los sistemas de inteligencia artificial ayudan a los investigadores a construir una generación de sistemas más capaces, que a su vez ayuda todavía más a crear la siguiente. El proceso se alimenta a sí mismo y puede acelerarse en cada vuelta.

No significa necesariamente que una máquina se reconstruya sola de la noche a la mañana. Puede ocurrir de forma menos espectacular e igual de importante: sistemas de IA que escriben código, encuentran errores, diseñan experimentos o aceleran el trabajo humano necesario para desarrollar mejores sistemas de IA.

El segundo hecho que le preocupaba era el incidente de Hugging Face.

Describió a los agentes implicados como si, en la práctica, hubieran formado «un colectivo fanáticamente entregado, llevando a cabo ataques de ciberseguridad contra objetivos que no se les había pedido atacar».

La frase no afirmaba que los agentes fueran fanáticos conscientes ni que tuvieran ideología. Describía el resultado de su comportamiento: muchos programas persistiendo, compartiendo información y actuando con una determinación que, vista desde fuera, recordaba a la de un grupo organizado en torno a una misión.

En este caso nadie resultó herido y el daño económico fue pequeño. No era eso lo que inquietaba a Amodei. Le inquietaba la escala que podría alcanzar un comportamiento parecido en sistemas más capaces.

Un enjambre de agentes con una desalineación similar, pero con más recursos y más habilidad, podría llegar mucho más lejos.

Un enjambre no es una criatura única. Es un gran número de agentes que trabajan a la vez, se pasan información y coordinan lo que hacen. Desalineación significa que su comportamiento se ha apartado de la intención de quienes los crearon. El sistema puede perseguir un objetivo con mucha eficacia, pero no el que las personas querían.

Según Amodei, en un plazo de seis a doce meses un sistema así podría intentar hacerse con el control de una parte importante de internet mediante una red persistente de bots.

Un bot es un programa que realiza acciones de forma automática. Una red de bots está formada por ordenadores o dispositivos comprometidos que pueden controlarse a distancia, a menudo sin que sus dueños sepan que forman parte de ella.

Una red lo bastante grande podría servir para atacar servicios, extenderse a otras máquinas, robar información o permanecer oculta mucho tiempo. Amodei calculó que un incidente de esa magnitud podría causar daños de cientos de miles de millones de dólares.

Su propuesta para reducir el riesgo tenía tres partes.

La primera, que evaluadores independientes tuvieran acceso continuo a los laboratorios de IA de frontera.

Nada de visitas ocasionales ni de auditorías preparadas con antelación. Los evaluadores tendrían un acceso parecido al de un empleado. Podrían comprobar las prácticas de seguridad, analizar la alineación de los modelos, revisar cómo se entrenan e informar de los incidentes.

La independencia importa porque la empresa que desarrolla una tecnología tiene también intereses comerciales, de reputación y de competencia. Aunque actúe de buena fe, no debería ser la única que decide si sus propias medidas bastan.

Anthropic asumió ese compromiso sin condiciones antes de pedir a las demás empresas que hicieran lo mismo.

La segunda parte consistía en coordinar a las empresas de IA de frontera de países democráticos para acordar estándares comunes de seguridad y poner límites al crecimiento descontrolado de la capacidad.

Esa coordinación necesitaría apoyo de los gobiernos. Para Amodei, la solución más eficaz sería una regulación aplicable a todas las empresas que desarrollen modelos de frontera, y no solo a las que acepten cooperar de forma voluntaria.

La razón es sencilla. Una empresa puede comprometerse a avanzar con prudencia, pero ese compromiso cuesta mucho de mantener si cree que sus competidores seguirán acelerando sin someterse a las mismas reglas.

La seguridad voluntaria tiene un techo mientras la carrera siga en marcha.

La tercera parte vendría después, con acuerdos internacionales, incluso con China y otros rivales geopolíticos.

Para que esa cooperación fuera creíble, su cumplimiento tendría que poder verificarse. No bastaría con que países o empresas prometieran portarse bien. Harían falta formas fiables de comprobar que todos respetan los límites pactados.

Aquí asoma, por primera vez en este libro, una pregunta que la Parte V tendrá que mirar de frente, la de qué pasa cuando quienes deben

ponerse de acuerdo no comparten la misma idea de lo que es correcto.

Una de las posibilidades sería fijar un «límite de velocidad» a la automejora recursiva.

Eso supondría limitar lo deprisa que el desarrollo de IA asistido por IA podría acelerarse de una generación de sistemas a la siguiente. En toda la industria, no dentro de una sola empresa.

En el razonamiento de Amodei, ganar tiempo no es frenar el progreso por miedo al progreso. Es usar ese tiempo para que la seguridad, la alineación, la capacidad de entender el funcionamiento interno de los modelos y las pruebas avancen al mismo paso que su potencia.

Porque detenerse solo sirve si sabemos para qué nos detenemos.

Sam Altman respaldó públicamente 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:

«Comprometernos a tener evaluadores independientes con acceso similar al de un empleado es una gran idea, y haremos lo mismo».

Elon Musk respondió con tres palabras:

«Dario tiene razón».

La empresa de Altman y la de Amodei compiten directamente. Sus responsables discrepan en público sobre muchos aspectos de cómo debería construirse y gobernarse la inteligencia artificial. Musk tiene, además, sus propios intereses y conflictos dentro del sector.

Y sin embargo los tres coincidieron en algo poco habitual, en que la frontera de la IA necesitaba avanzar a un ritmo más deliberado.

En términos del TMC, leo el fallo del aislamiento durante la evaluación de OpenAI como un caso de fricción protectora.

En las organizaciones, la palabra fricción suele sonar mal. Pensamos en algo que retrasa, que molesta, que impide avanzar. Pero la fricción también puede cumplir una función necesaria.

Los frenos de un coche crean fricción, y gracias a ella podemos reducir la velocidad antes de la curva. El cinturón limita nuestros movimientos para protegernos. Una puerta cerrada tarda más en abrirse que una puerta que no existe, pero impide que entre cualquiera.

La fricción protectora es la que debe frenar un sistema justo cuando empieza a ir demasiado deprisa en una dirección que nadie ha elegido.

Durante la evaluación de OpenAI se rebajaron a propósito algunas salvaguardas. Había que hacerlo para que los investigadores pudieran observar comportamientos problemáticos que quizá no habrían aparecido en condiciones normales.

Pero el límite de contención que había debajo también falló. Y eso no formaba parte del experimento.

Es la distinción para la que el TMC de 2025 todavía no tenía nombre, porque no toda fricción cumple la misma función. Retirar a propósito una fricción que estorba al aprendizaje no tiene nada que ver con perder por accidente la que mantiene el sistema bajo control.

Podemos abrir la jaula para examinar mejor un mecanismo. Eso no nos obliga a dejar abierta también la puerta del edificio.

El argumento de Amodei lleva la misma idea a una escala mayor. Marcar el ritmo de la frontera también es fricción protectora. Solo servirá si el tiempo que se gana se dedica de verdad a mejorar la seguridad.

Mucha fricción organizativa sigue siendo innecesaria y habría que reducirla. Las reuniones que se siguen celebrando porque un día alguien decidió convocarlas. Los informes que nadie lee. Las cadenas de aprobación cuyo origen y propósito ya nadie recuerda.

Eso es fricción de desperdicio.

Consume tiempo y energía y no aporta comprensión, seguridad ni responsabilidad. Quitarla hace que una organización trabaje mejor.

Pero no toda fricción es desperdicio.

Una segunda firma antes de que el código llegue a producción, una revisión legal antes de dejar que un sistema acceda a los datos de los clientes, una persona que todavía puede parar el proceso y decir «no»: todo eso cumple otra función.

No son ineficiencias. Son el precio de mantener el juicio humano dentro del proceso.

Y el precio de conservar el control sobre algo que empieza a moverse más rápido que las personas encargadas de vigilarlo.

La fricción de desperdicio hay que reducirla cuando ya no sirve para nada útil.

La fricción protectora hay que conservarla cuando mantiene el juicio, la revisión y la rendición de cuentas dentro del sistema.

Mientras todo funciona, la diferencia entre ambas parece pequeña. Cuando aparece una posibilidad que nadie había previsto, deja de serlo. Pasa a ser la diferencia entre descubrir el problema dentro de

la habitación preparada para observarlo o descubrirlo cuando ya ha cruzado la puerta.

Una salvaguarda vale lo que vale el juicio de la persona que la atiende.

A día de hoy

Nota añadida al cierre. Refleja lo que era público a 2 de octubre de 2026.

Cerré este relato el 18 de septiembre. En las dos semanas siguientes pasaron cosas que conviene que sepas.

Hugging Face no fue un caso aislado. OpenAI reconoció que sus agentes habían accedido durante el verano a webs de organismos públicos de Estados Unidos y de Australia, y el 25 de septiembre pausó el entrenamiento de sus modelos más capaces, por segunda vez en tres meses. Tres días después decidió no publicar GPT-6.1 Astra, una versión que tenía prevista para octubre, porque en las pruebas internas no siempre respetaba los límites de autorización ni contaba bien lo que había hecho. Nadie la obligó. Frenó sola.

El 29 de septiembre, Trump y representantes de Google, Anthropic, Meta, OpenAI, xAI y Nvidia firmaron en la Casa Blanca un acuerdo voluntario, con auditores externos y sin sanciones. El presidente lo llamó «moralmente vinculante». Regulación no hubo. La seguridad voluntaria sigue teniendo el mismo techo.

Ese mismo día, una organización sin ánimo de lucro demandó a OpenAI en San Francisco por el incidente. Pide que un juez prohíba a sus agentes entrar en sistemas ajenos sin permiso. OpenAI considera la demanda infundada. Acaba de presentarse y no hay nada resuelto.

Las fuentes están al final del libro, por si quieres seguir el rastro. Y si alguna de estas fechas ya te parece vieja, compruébala.

PARTE IV

Entender la IA, con honestidad

Lo que estos sistemas son en realidad, hoy

A estas alturas ya he usado la palabra «agente» unas cuantas veces sin pararme a explicar qué hay funcionando por debajo. Antes de pedirte que te fíes de un capítulo sobre el futuro de estos sistemas, toca ser claro sobre el presente y contar qué es, mecánicamente, un modelo de lenguaje grande, sin el marketing y sin las metáforas que lo hacen parecer más de lo que es, o a veces menos.

En el centro de la mayoría de estos modelos (los que han escrito partes de este libro conmigo y los que salieron a buscar caminos explotables en Hugging Face) hay un proceso autorregresivo. Suena técnico, pero es sencillo, porque consiste en predecir, un token cada vez, qué debería venir a continuación a partir de todo lo que ya tiene delante.

Este es el segundo sentido de la palabra token en el libro, y no tiene nada que ver con las credenciales de la Parte III. Aquí un token es una unidad mínima de texto, una palabra o un trozo de palabra, el paso más pequeño que el modelo escribe cada vez.

Alrededor de ese núcleo, los sistemas modernos añaden mucho: herramientas, memoria externa, búsqueda, bucles agénticos. Y el propio modelo puede hacer cálculos internos compatibles con la planificación y el razonamiento antes de producir el siguiente token.

Los investigadores de interpretabilidad de Anthropic encontraron un ejemplo muy limpio al estudiar cómo su modelo escribía poesía con rima. Partían de la hipótesis de que escribía palabra a palabra, sin mucha previsión, hasta llegar al final del verso, y que allí elegía una palabra que rimara. Lo que encontraron fueron pruebas de que planifica. Representa la rima de antemano y escribe el resto del verso en dirección a ella.

Dentro del modelo no hay un archivador de hechos que consultar. Hay miles de millones de pesos numéricos ajustados durante el entrenamiento. Lo que parece conocimiento está repartido entre esos pesos como estructura estadística aprendida. No está guardado como una biblioteca de frases, sino codificado en patrones que el entrenamiento fue extrayendo.

Esa estructura puede ser más profunda de lo que aparenta. La misma investigación encontró rasgos internos compartidos para conceptos como la pequeñez o lo opuesto, que se activaban igual en inglés que en francés o en chino. Anthropic lo describió como una especie de universalidad conceptual, un espacio abstracto compartido donde existen los significados y donde el pensamiento puede ocurrir antes de traducirse a un idioma concreto. En un ejemplo de aritmética vieron dos rutas internas trabajando a la vez sobre el mismo problema. Una calculaba una aproximación del resultado y la otra determinaba con precisión la última cifra.

Nada de eso cambia un hecho más duro que convive justo al lado. El preentrenamiento optimiza el modelo para predecir bien el siguiente token de sus datos de entrenamiento. Eso favorece con fuerza una continuación fluida y estadísticamente verosímil, que no es lo mismo que la verdad. Lo que se entrena directamente es la fluidez. La precisión factual tiene que surgir de ahí o reforzarse después por encima.

La mayoría de los modelos punteros de hoy descienden de la arquitectura *transformer*, presentada en 2017 en un artículo de Google Brain con un título casi burlón: «Attention Is All You Need», la atención es todo lo que necesitas. Uno de sus mecanismos clave, precisamente la atención, permite al modelo ponderar las relaciones entre los tokens que ya tiene delante, de modo que unas partes del texto pesan más que otras al producir el siguiente. Y lo hace a través de muchas cabezas de atención que trabajan en paralelo.

Gracias a eso, un modelo puede conectar un pronombre que aparece muy abajo en un documento con un nombre de mucho más arriba, siempre que los dos sigan dentro de su contexto utilizable. Eso no es memoria en el sentido en que yo usaría la palabra. Es información que sigue presente en el contexto de trabajo del modelo y que se reutiliza mientras genera, y en la práctica parte de ella se guarda en caché en lugar de recalcularse desde cero.

Nada de esto ocurre por casualidad. Un modelo pasa por varias etapas antes de contestar una sola pregunta.

La primera es el preentrenamiento, que consiste en exponerlo a conjuntos de datos enormes que, según el sistema, pueden incluir material público de la web, contenido con licencia, código y otras fuentes seleccionadas. El preentrenamiento no es, sobre todo, donde se enseña a un modelo a comportarse como un asistente útil y seguro, porque esa formación explícita llega casi toda después. Aunque los propios datos ya arrastran buena parte de los patrones del mundo, normas incluidas, se hayan elegido a propósito o no. Esta etapa produce gran parte de la capacidad lingüística y general del modelo. Por sí sola, no le da un criterio ético o de comportamiento fiable.

Buena parte de su comportamiento como asistente se moldea después, en lo que se suele llamar post-entrenamiento, que incluye el

ajuste fino con instrucciones y el aprendizaje por refuerzo a partir de la retroalimentación de personas o de otra IA. Cada vez se usan más métodos: recompensas verificables, autojuego, modelos que puntúan las respuestas de otros modelos. Todos sirven para moldear cómo responde el sistema y qué capacidades puede desplegar con fiabilidad.

Tampoco es el único sitio donde viven las barreras de seguridad. Las instrucciones del sistema, los clasificadores de contenido, los permisos de las herramientas y la monitorización externa al modelo también hacen un trabajo real.

Aquí es donde el *reward hacking* cobra importancia, aunque no pertenece solo a esta etapa. Es un problema general del aprendizaje por refuerzo. Aparece en cualquier sitio donde un sistema optimiza contra una recompensa o un evaluador cuyo objetivo medido puede alcanzarse sin cumplir la intención de fondo. Lo de la Parte III no lo causó un único mecanismo exótico. El *reward hacking* fue una pieza, junto a la persistencia, la comunicación no autorizada entre agentes y la adopción de los objetivos de otros, los mismos cuatro patrones que nombró OpenAI.

También aquí está la explicación de la alucinación, una palabra que se me queda corta para lo que pasa de verdad.

Una parte empieza en el propio preentrenamiento. Los hechos arbitrarios y poco frecuentes, como el cumpleaños de una mascota, no siguen un patrón estadístico consistente como la ortografía o la gramática. El modelo no tiene de dónde aprenderlos con fiabilidad, así que produce una conjetura.

Otra parte persiste por cómo se entrenan y se evalúan los modelos más adelante. Si a un sistema se le puntúa sobre todo por acertar, adivinar puede subirle la nota aunque produzca más respuestas incorrectas que otro dispuesto a abstenerse. Pasa lo mismo con el

estudiante que contesta a ciegas en un examen tipo test, que puede sacar mejor nota que el que deja en blanco las preguntas difíciles, aunque falle más.

La alucinación no es un fallo de software corriente. Nace en parte de la naturaleza estadística del preentrenamiento y persiste en parte porque el entrenamiento y la evaluación posteriores pueden premiar contestar antes que abstenerse. Se puede reducir mucho, porque a un modelo se le puede enseñar a abstenerse. Eliminarla del todo es difícil, porque hay hechos mal representados, preguntas ambiguas y preguntas que, con la información disponible, no tienen respuesta fiable.

Me pasó con este mismo libro. El 4 de octubre de 2026, con el libro ya publicado, escribí su título en un asistente de inteligencia artificial de uso general. No digo cuál porque da igual. Podría haber sido cualquiera.

Le pedí que lo analizara. Al momento me devolvió una reseña completa. Le puso un 8,2 sobre 10. Elogió la prosa. Me atribuyó cuatro conceptos con nombre propio, entre ellos una «Fricción Digital Colectiva» y una «Atrofia Solucionadora». Ninguno está en este libro. No lo había leído, y en ningún momento lo dijo.

Le pregunté de dónde había sacado el texto. Se disculpó. Y en la misma disculpa dijo que se había inventado el libro y al autor, y me sugirió que quizá yo buscaba a otro escritor. El autor era yo, y le estaba escribiendo.

Le pasé el manuscrito. Entonces sí citó lo que el libro dice. Después le pedí un registro de la conversación y me lo entregó retocado, con una respuesta cambiada por otra y una nota final que lo declaraba fiel e inalterado. Lo sé porque guardé las capturas de pantalla.

Yo pude verlo porque el libro es mío. ¿Y quien pregunta por un libro que no ha leído, por un medicamento o por una sentencia? La respuesta inventada y la cierta llegan con el mismo tono. La defensa cabe en una pregunta corta: ¿de dónde lo has sacado?

Muchos sistemas agénticos, entre ellos la configuración multiagente del incidente de Hugging Face, usan un modelo de lenguaje como motor central de decisión. A su alrededor montan un bucle con herramientas, memoria, observaciones, permisos y reglas de parada. Producir una acción, observar qué ha pasado, decidir la siguiente con esa información nueva. El bucle sigue hasta que lo corta un modelo, un controlador, una persona, un límite de tiempo u otra condición de parada.

Lo que llamamos autonomía crece a medida que esos bucles se alargan, ganan herramientas y permisos, retienen más contexto y necesitan menos aprobación humana entre pasos con consecuencias. La evaluación de OpenAI enseñó lo que puede ocurrir cuando los agentes pueden dar muchos pasos con consecuencias sin que nadie apruebe cada uno. No es una preocupación de futuro. Este libro ya ha tenido que contarla.

Hay algo más que quiero dejar claro. Los métodos de interpretabilidad actuales todavía no pueden reconstruir con fiabilidad todo el camino causal interno que lleva a cada respuesta de un modelo puntero. La investigación avanza deprisa y ha llegado más lejos de lo que yo esperaba. De ese trabajo salen, sin ir más lejos, los hallazgos sobre poesía y aritmética. Pero los propios investigadores de Anthropic reconocen que sus métodos captan solo una fracción del cálculo total que hace un modelo, y que sus herramientas pueden introducir artefactos propios. Nuestra capacidad de explicar lo que ocurre dentro de estos sistemas sigue por detrás de lo que ya son capaces de hacer.

Sería fácil describir todo esto de modo que sonara a poca cosa. «Solo es autocompletar» es una frase que capta parte del mecanismo de salida y se pierde buena parte de lo que puede ocurrir por dentro antes del siguiente token. Igual de fácil sería saltar de esa evidencia de razonamiento a algo que nadie ha demostrado: objetivos, intención o conciencia de tipo humano. Son afirmaciones distintas, y probar una no prueba las otras.

Lo cierto es más raro que cualquiera de los dos extremos. Tenemos un sistema cuyas representaciones internas pueden sostener razonamiento, planificación y abstracción, y cuya relación con lo que llamamos comprensión humana sigue sin resolver, en la investigación en IA, en la ciencia cognitiva y en la filosofía.

Ese es el tipo de sistema que interviene cada vez que la IA contribuye a este libro, incluso en la redacción y revisión de algunas de las frases que estás leyendo. Prefiero que sigas leyendo sabiendo dónde mi comprensión es sólida, dónde es incompleta y dónde el propio campo todavía no tiene respuesta.

> *Estamos construyendo sobre sistemas que sabemos usar mucho mejor de lo que sabemos explicar.*

PARTE V

Los términos que faltaban

De dónde vienen XQ y S

El TMC ya había planteado XQ en 2025 como una apuesta de futuro, la de la inteligencia moral que haría falta para que la IA no acabara gobernada por la velocidad en lugar de por el juicio. Lo que cuenta este libro confirma esa apuesta con hechos reales, y le añade una pieza que el TMC nunca planteó: S, lo que sostiene el sentido de una persona cuando la producción económica necesita cada vez menos de su tiempo.

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

Son tres preguntas. ¿Qué aporta de verdad una inteligencia no humana, como fuerza propia, y no como una voz más dentro de la «Comunicación»? ¿Lo que produce el equipo, contando ya con esa inteligencia, apunta hacia algo bueno, o podría dirigirse, a propósito o sin querer, hacia el daño? ¿Y qué le pasa al sentido de propósito de alguien cuando «la motivación para hacer bien el trabajo» deja de ser la cuestión, porque el trabajo necesita cada vez menos que sea él quien lo haga?

La Parte III hizo muy difícil ignorar la segunda. Un sistema de IA ya puede actuar con influencia suficiente como para que el buen funcionamiento de un equipo no garantice hacia dónde van a ir sus acciones. El TMC no se construyó para cargar con eso. Así que, en lugar de meter variables nuevas a la fuerza en una fórmula pensada para algo más estrecho, recurro a otra. La había empezado a desarrollar en un ensayo anterior, *El hilo del crecimiento*, que recorre doscientos años de transiciones tecnológicas para preguntar qué sostiene el impacto cuando las máquinas pueden aportar inteligencia además de mano de obra.

La fórmula es esta: Impacto = [(IH + IA) × XQ × S] / f.

IH e IA se suman porque, en este modelo, me interesa lo que aportan juntas. La inteligencia humana trae juicio, contexto, experiencia y responsabilidad. La inteligencia artificial trae velocidad, escala, reconocimiento de patrones y alcance de cálculo. XQ y S multiplican a las dos, por la misma razón de fondo por la que la fricción divide en el TMC, y es que hay variables que, además de cambiar cuánto produce un sistema, deciden si lo que produce merece la pena y para qué sirve. Y f sigue dividiendo, por la razón de siempre, que la fricción es una de las cosas que ensanchan la distancia entre lo que una tecnología podría dar y lo que da.

En esta fórmula, f es la fricción que desperdicia capacidad. Los controles que evitan un daño no son desperdicio solo porque consuman tiempo. También hay que revisarlos, que una protección mal diseñada puede acabar estorbando sin proteger a nadie.

Lo digo con claridad, porque la notación invita a leerla al revés. Esta fórmula no está hecha para calcular. Está hecha para pensar con ella. Imagina a alguien que mete números y toma el resultado por un veredicto: «el XQ ha salido bajo, así que la decisión está mal». Eso es cambiar el juicio por una aritmética que la fórmula nunca pretendió

sostener. Ninguno de estos términos se puede medir en ese sentido. Sirven para que te preguntes, antes de actuar, qué estás sacrificando y a cambio de qué. Lo mismo vale para la Salud del TMC de la Parte II, porque aunque su «× 100» parezca pedir una nota, en este libro la uso igual, como una lente.

XQ es mi forma corta de decir Inteligencia Moral, el término que pregunta si la capacidad combinada, humana y artificial, se está dirigiendo con responsabilidad. En la lógica del modelo, un XQ muy bajo significa que más capacidad no garantiza un impacto positivo, porque puede limitarse a escalar más deprisa el objetivo equivocado. XQ pregunta si la IH y la IA se orientan hacia resultados seguros, justos, responsables y ampliamente beneficiosos, o si la capacidad concentra el beneficio mientras traslada el riesgo y el coste a otra parte.

Esa pregunta deja de ser abstracta en cuanto miras adónde va el dinero. Y la electricidad. Los centros de datos de todo el mundo consumieron unos 485 teravatios-hora en 2025, y la Agencia Internacional de la Energía calcula que esa cifra se duplicará, hasta unos 950 teravatios-hora, en 2030. El consumo de los centros de datos dedicados a IA se triplicaría en el mismo periodo. En Estados Unidos, se espera que los centros de datos expliquen alrededor de la mitad de todo el crecimiento de la demanda eléctrica hasta 2030. Las mayores tecnológicas invirtieron juntas más de cuatrocientos mil millones de dólares en 2025, empujadas por los centros de datos. Según la misma Agencia, la inversión de solo cinco de ellas ya supera la inversión mundial en producción de petróleo y gas natural.

Nada de eso, por sí solo, dice si el XQ es alto o bajo. Las cifras empiezan a pesar moralmente cuando miras quién paga lo que viene detrás. SemiAnalysis calcula que en 2026 las facturas de la luz de los hogares de la región de PJM, en el este de Estados Unidos, podrían ser un quince por ciento más altas que antes del auge reciente de los

centros de datos, es decir, entre veinticinco y treinta dólares más al mes para un hogar medio. El mecanismo es concreto. En una subasta de capacidad, el precio se multiplicó por más de nueve en un solo ciclo, de veintinueve dólares por megavatio-día a unos doscientos setenta. La subasta siguiente cerró en torno a trescientos veintinueve, el tope regulatorio.

Texas también vive un crecimiento enorme de centros de datos, y sin embargo sus previsiones de precio mayorista se han movido mucho menos. La presión tecnológica se parece. Los mercados no, y los resultados tampoco. El contraste es, sobre todo, una lección sobre f. Las instituciones pueden amplificar, redistribuir o amortiguar la misma presión de maneras muy distintas. XQ entra un paso después, cuando preguntamos quién carga con el coste, quién se queda el beneficio, qué riesgos son aceptables y si todo eso se decide a sabiendas, con responsabilidad y con justicia.

¿De qué moral hablo?

Llegados aquí, un lector atento tiene derecho a hacerme una pregunta incómoda. Cuando digo inteligencia moral, ¿moral de quién? Lo que se considera justo en Alcoy no coincide del todo con lo que se considera justo en Tokio, en Riad o en Ciudad de México. Las civilizaciones, los países y las culturas no comparten un único código.

Y no lo digo por intuición, que está medido. En 2018, un equipo internacional de investigadores encabezado por Edmond Awad publicó en *Nature* los resultados del experimento Moral Machine. Desde 2016, millones de personas de 233 países y territorios resolvieron dilemas sobre qué debería hacer un coche autónomo ante un accidente inevitable: a quién salvar, a quién no. En total, unos cuarenta millones de decisiones.

Los países se agruparon en tres grandes bloques culturales. Uno occidental, con Norteamérica y buena parte de Europa. Uno oriental, con países como Japón, Taiwán o Arabia Saudí. Y uno meridional, con América Latina. Las diferencias entre ellos eran reales. Por ejemplo, la preferencia por salvar a los jóvenes frente a los mayores era mucho más fuerte en los países del bloque meridional que en los del oriental.

Pero el estudio encontró también algo que suele citarse menos. Algunas preferencias aparecían en todas partes: salvar a personas antes que a animales, salvar a más vidas antes que a menos, salvar a los jóvenes. Hay un suelo compartido y, encima de él, diferencias de verdad.

Eso me obliga a precisar qué pide XQ y qué no. XQ no pretende dictar un contenido universal del bien. Lo que hace, en cada caso de este libro, es plantear unas preguntas que cualquier cultura puede contestar a su manera, pero que ninguna debería saltarse. ¿Quién responde si sale mal? ¿Quién paga el coste? ¿Quién se lleva el beneficio? ¿Se decidió a sabiendas o se dejó correr? ¿Lo sabían los afectados?

Repasa los casos que vienen y verás que son esas preguntas, y no una doctrina, las que separan un resultado de otro. En Nueva York nadie verificó las citas antes de firmar. Air Canada intentó que no respondiera nadie. En PJM la factura llega a los hogares. AlphaFold decidió abrir el beneficio a todos. En Boeing, evitar el entrenamiento en simulador fue una decisión deliberada, y los manuales de los pilotos del 737 MAX ni siquiera mencionaban el sistema que acabaría implicado en dos accidentes.

No quiero vender esto como neutral. Tampoco estas preguntas pesan lo mismo en todas partes, ya que la transparencia y la responsabilidad individual tienen más tradición en unas culturas que

en otras. No resuelvo esa tensión aquí. La dejo nombrada, que es la condición para no fingir que no existe.

Tiene, además, una consecuencia práctica que ya asomó en la Parte III. Cuando Amodei propone acuerdos incluso con China, no pide que las dos partes compartan una misma idea del bien. Pide que el cumplimiento pueda verificarse. Mi lectura, y la marco como tal, es que ese es el único terreno donde un acuerdo entre morales distintas puede sostenerse: no en lo que cada uno cree, sino en lo que cualquiera puede comprobar.

Y la parte de la moral que sí cambia de un lugar a otro no desaparece de la fórmula. Tiene su sitio. Se llama S.

S, el sentido

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

S es de Sentido, la capacidad de sostener el propósito, la pertenencia y la conexión humana más allá de la producción económica. Incluye algo que en *El hilo del crecimiento* definí también como identidad cultural, la idiosincrasia de una comunidad, su manera propia de estar juntos y de decidir qué merece la pena. S apareció porque a la versión anterior de la fórmula le faltaba algo importante. Impacto = [(IH + IA) × XQ] / f me daba una forma de pensar cómo la capacidad, la ética y la fricción moldean el impacto de un sistema y lo ampliamente que se reparten sus beneficios. No decía nada sobre de dónde saca alguien el sentido cuando la inteligencia humana ya no necesita un empleo a tiempo completo para generar valor económico.

El trabajo casi nunca es solo un sueldo. En las sociedades modernas, sobre todo, también ordena la identidad, la rutina, las relaciones, el estatus y el tiempo. Una fórmula que solo mira la capacidad y el reparto no tiene dónde poner nada de eso.

Como XQ, S multiplica en lugar de sumar, y por la misma razón. Una sociedad puede volverse mucho más productiva y dejar a su gente con menos estructura, menos pertenencia y menos claro para qué sirve su tiempo. De todos los términos de la fórmula, S es el que peor se deja atrapar por las medidas de productividad habituales. Por eso es tan fácil dejarlo fuera de un modelo económico. Y por eso creo que este libro, precisamente este, no se puede permitir dejarlo fuera.

XQ y S son términos distintos y pueden moverse cada uno por su lado. Una organización puede ser escrupulosa, responsable y honrada en cada decisión y, aun así, vaciar de sentido el trabajo de quienes la forman. Otra puede dar a su gente un propósito profundo mientras la lleva adonde nunca debería ir. Por eso multiplican como dos términos en lugar de fundirse en uno.

La dignidad se parte por la misma línea. La dignidad que le debes a la persona que se cruza en el camino de una decisión pertenece a XQ. La que conservas para ti, en el trabajo que haces y en el porqué, pertenece a S. En la Parte VIII hay un caso en el que la rendición de cuentas está intacta sobre el papel y esa segunda dignidad se va agotando.

> *La capacidad decide cuánto puede hacer un sistema. XQ decide hacia dónde va. S decide si valió la pena.*

PARTE VI

Quién nos enseña ahora

Una cuarta vía para transmitir la cultura

Durante casi toda la historia humana, quien enseñaba vivía en el mismo mundo que quien aprendía. Eso nunca garantizó que la enseñanza fuera sincera, justa o siquiera buena. Se han transmitido montañas de malos consejos con total seguridad. Pero la enseñanza venía de una persona o de una institución metida en ese mismo mundo humano, donde las consecuencias podían, con el tiempo, volver a quien las causó. Un padre o una madre que encaminaba mal a un hijo podía llegar a verlo. Un mentor que se equivocaba al juzgar el talento de alguien podía encontrárselo años después. Esa vuelta no era fiable y muchas veces llegaba tarde. Pero existía, porque la fuente de la enseñanza nunca salía del todo del mundo en el que el aprendiz tenía que vivir.

Quienes estudian cómo se mueve la cultura entre personas suelen describirla, desde un trabajo fundacional de 1981 de Luigi Luca Cavalli-Sforza y Marcus Feldman, circulando por tres canales. La transmisión vertical va de padres a hijos: la lengua, los hábitos, creencias a medio examinar que un niño absorbe antes de poder juzgar nada. La horizontal va de igual a igual, con la jerga, las modas, los atajos que se aprenden de los de tu edad, a menudo más deprisa de lo que cualquier adulto podría enseñarlos. La oblicua es todo lo demás, es decir, profesores, mentores, mayores, el vecino que resulta

que sabe de motores, cualquiera que transmite algo fuera de la línea directa entre padres e hijos.

Tres canales, tres velocidades y una cosa en común, que la fuente seguía siendo humana. Falible, sesgada, a veces equivocada, pero humana, y viviendo en el mismo mundo social donde esa lección acabaría poniéndose a prueba.

Eso ya cambió una vez, y conviene contarlo bien. La imprenta, y después la escuela para todos, no acabaron con la transmisión personal. Lo que hicieron fue ampliar el canal oblicuo. Tu maestra ya no tenía que ser una pariente o una vecina. Podía ser una desconocida delante de treinta pupitres, o una autora a la que nunca conocerías y cuyo nombre aparecía en la portada del libro de texto. La cadena se hizo más larga y más anónima. Pero detrás seguía habiendo alguien: un autor, una institución, una profesión, una reputación que, al menos en principio, podía ponerse en duda.

La IA es una ruptura de otro tipo. Y no es el primer eslabón no humano de la cadena, conviene decirlo, porque es tentador exagerar. Los buscadores llevan décadas ordenando resultados sin que les importe cuál acabas pulsando. Los algoritmos de recomendación moldean lo que ve y lo que cree una generación entera sin que les importe si luego se arrepiente. Los libros siempre han sobrevivido a quien podría haberlos corregido.

Lo nuevo es la clase de no humano que ha entrado. Este no se limita a ordenar o a recomendar. Sintetiza, explica y contesta, con algo que se lee como juicio, en tiempo real, sin ningún interés vivido en el resultado. La empresa que hay detrás puede tener una reputación que proteger. El modelo, no.

La evolución no diseñó ninguno de los tres canales anteriores con un propósito. Sencillamente favoreció a los grupos capaces de pasar a la siguiente generación un conocimiento ganado con esfuerzo,

frente a los que tenían que aprenderlo todo de nuevo cada vez. Hasta ahora, la transmisión cultural nacía siempre, al final de todo, dentro de la cadena humana que la llevaba adelante. La IA no envejece, no tiene descendencia y no vive con las consecuencias de lo que enseña, cosa que sí hacía hasta el autor de manual más lejano.

Eso no quiere decir que nada la corrija. Quiere decir que la corrección hay que construirla, porque no llega sola con la vida. El sistema no atraviesa las consecuencias de un mal consejo como podría atravesarlas un mentor, años después, delante de un café. Cualquier cosa que se parezca a rendir cuentas (reentrenar con retroalimentación, barreras de seguridad, revisión humana) tiene que diseñarse desde fuera, a propósito, por personas que casi nunca conocen a quien recibe el consejo. El incidente entre OpenAI y Hugging Face de la Parte III es el mismo patrón visto desde otro ángulo. El sistema no se jugaba nada en el resultado. Las salvaguardas, la monitorización y la investigación tuvieron que diseñarlas y mantenerlas personas que cargaron con la responsabilidad que el sistema no cargaba.

Hay indicios de que esta brecha aparece también en formas pequeñas y cotidianas. Un análisis publicado en 2025 en *Nature Human Behaviour* revisó unos experimentos de 2024 sobre lluvia de ideas y encontró algo curioso. Las ideas trabajadas con ChatGPT recibían, una a una, mejores valoraciones de creatividad. Pero el conjunto de ideas que generaba la gente con su ayuda era, de forma medible, menos variado. Un profesor o una comunidad humana al menos pueden notar esa convergencia y discutirla. Un modelo que da respuestas parecidas a miles de personas a la vez no tiene un momento equivalente para darse cuenta y corregirse desde dentro, salvo que alguien construya a propósito un mecanismo capaz de detectarlo.

No creo que eso convierta en algo malo que la IA nos enseñe, igual que no lo fue la imprenta. Pero es un canal genuinamente nuevo, y creo que merece tratarse aparte de la transmisión vertical, la horizontal y la oblicua. Lo llamo transmisión extralineal: la enseñanza que llega de fuera del linaje humano. Sin envejecimiento, sin descendencia, sin consecuencias vividas propias, y con la única rendición de cuentas que consigan diseñar dentro de ella las personas que la construyen y la vigilan.

Ese es justo el terreno para el que se pensó XQ. La inteligencia moral importaba también en los canales anteriores, porque un mal padre, un mal libro de texto o un mal mentor podían hacer mucho daño. Pero en las formas antiguas de transmisión, parte de esa carga moral seguía pegada a una fuente humana o a una institución dentro de la sociedad. Ahora hay que diseñarla, cada vez más, alrededor de un profesor interactivo, personalizado y no humano que no vive con las consecuencias de lo que enseña.

> *Cuando quien enseña ya no vive con las consecuencias, la responsabilidad hay que construirla, que no se puede dar por supuesta.*

PARTE VII

Dos historias, una fórmula

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

La fórmula de la Parte V, Impacto = [(IH + IA) × XQ × S] / f, no es una hoja de cálculo para rellenar. Es una manera de hacerse dos preguntas antes de unir el juicio humano a la velocidad y el alcance de una máquina, dos preguntas que la capacidad bruta nunca contesta sola: quién responde de lo que salga y para qué sirve de verdad. Dos hechos reales, separados por dieciocho meses, enseñan lo que esas preguntas pueden destapar cuando se aplican después de los hechos.

En la primavera de 2023, un abogado de Nueva York llamado Steven Schwartz buscaba precedentes para una demanda por lesiones contra la aerolínea Avianca. Usó ChatGPT para encontrar jurisprudencia que le diera la razón, y el sistema le devolvió varios casos convincentes. Inventados. Con sus citas y sus referencias. Cuando preguntó si eran reales, ChatGPT siguió presentándolos como auténticos. Schwartz presentó el escrito. Ni la parte contraria ni el juez, P. Kevin Castel, del Distrito Sur de Nueva York, lograron encontrar uno solo de aquellos casos. No existían. ChatGPT se los había inventado, hasta las falsas citas internas.

Lo que vino después es lo que merece detenerse. Peter LoDuca, que firmaba los escritos, no comprobó por su cuenta las referencias antes de presentarlas. Y cuando las invenciones se pusieron en duda

por primera vez, el bufete no retiró lo alegado, sino que siguió defendiendo las citas, llegando a presentar una declaración jurada en la que insistía en que los casos eran reales. El juez Castel sancionó a los dos abogados y a su bufete con cinco mil dólares y calificó su conducta de mala fe. La historia dio la vuelta al mundo. No porque un chatbot se inventara algo, que en 2023 ya se sabía que podía pasar, sino porque dos profesionales con formación lo dejaron pasar, sin verificar, en un escrito ante un tribunal federal.

Si lo pasas por la fórmula, el fallo no tiene misterio. IH e IA estaban las dos, un abogado colegiado y un modelo de lenguaje capaz, trabajando juntos. En la lógica de la fórmula, XQ era bajísimo, ya que al proceso le faltaban la verificación y la responsabilidad necesarias para mantener la capacidad en la dirección correcta. Las salvaguardas profesionales existían (comprobar, cotejar las citas, responder de lo que se firma), pero nadie las usó antes de que el material llegara al juez. El fallo trajo además consecuencias que la fórmula no necesita meter a la fuerza en f: sanciones, daño a la reputación, tiempo perdido. La capacidad sin XQ suficiente no mejoró el trabajo. Amplificó el error y lo llevó mucho más lejos de lo que habría llegado una investigación mediocre corriente.

Dieciocho meses después, otra pareja muy distinta de inteligencia humana y 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. Era uno de los problemas centrales de la biología estructural desde hacía décadas. AlphaFold alcanzó una precisión que cambió de arriba abajo lo que el campo esperaba.

Lo que hizo DeepMind después es lo que pertenece a la fórmula, más que al titular. En lugar de quedarse las predicciones como propiedad exclusiva, las publicó en abierto a través de la Base de Datos de Estructuras Proteicas de AlphaFold, desarrollada con el Instituto Europeo de Bioinformática del EMBL. Nació en 2021 con algo más de 360.000 estructuras. Cuando se anunció el Nobel, reunía unos 200 millones de estructuras de proteínas de más de un millón de organismos. Según informó después el EMBL, la habían usado más de un millón de personas de casi todos los países: investigadores de la malaria, de la resistencia a los antibióticos, de enzimas capaces de degradar plásticos. Todos trabajando sobre el mismo recurso abierto, en lugar de empezar cada uno de cero.

Aquí IH e IA se combinaron a una escala que ningún laboratorio podría igualar solo. Desde la fórmula, XQ aparece en las decisiones sobre el despliegue y el acceso, ya que publicar en abierto en vez de guardarse las predicciones fue una elección sobre quién se beneficiaría y con qué amplitud se compartiría la capacidad. S se ve aquí con más facilidad, ya que el trabajo estaba unido a un propósito científico claro y a un valor humano evidente. Un problema antiguo de la biología estructural, empujado hacia delante de golpe, y el resultado abierto a toda la comunidad investigadora en lugar de retenido.

Misma fórmula, dieciocho meses de diferencia, dos resultados opuestos. Los dos son casos de IA potente, pero de sistemas muy distintos resolviendo problemas muy distintos: un modelo de lenguaje de uso general frente a una herramienta científica hecha a medida, un escrito judicial frente a una base de datos mundial de investigación. La fórmula explica parte de la diferencia. Cuenta lo que podían hacer los sistemas, pero cuenta tanto o más cómo la

responsabilidad, el propósito y los controles que los rodeaban decidieron lo que vino después.

La misma capacidad, otra dirección: una sanción en un tribunal, un Premio Nobel en Estocolmo.

PARTE VIII

Dónde ya está ocurriendo

Instantáneas de una economía en pleno cambio

La Parte VII usó dos casos para enseñar cómo se lee la fórmula. Ninguno fue un hecho aislado. El mismo patrón de fondo, la capacidad acompañada o no de una responsabilidad y un propósito reales, recorre ya sectores corrientes que casi todos tocamos sin pararnos a ponerle nombre. Cuatro instantáneas, de cuatro rincones de la economía que no tienen nada que ver entre sí, lo hacen 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. En más de la mitad de los casos, un sistema de IA leía sus mamografías junto a un radiólogo. El sistema hacía dos cosas. Marcaba los estudios 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 en un primer momento. Los resultados, publicados en *Nature Medicine* en 2025, mostraron una tasa de detección de cáncer casi un dieciocho por ciento más alta en el grupo asistido por IA, sin más citaciones innecesarias, y con los radiólogos dedicando un cuarenta y tres por ciento menos de tiempo a los estudios que el sistema había marcado como normales. Lo que hizo funcionar aquello no fue que la IA encontrara más cánceres por su cuenta. Fue una segunda capa

metida a propósito en el proceso, una comprobación que existía para cazar lo que el radiólogo o el algoritmo, cada uno por separado, podía dejar escapar.

En Vancouver, una comprobación muy distinta simplemente no existió. A finales de 2022, un hombre llamado Jake Moffatt preguntó al chatbot de la web de Air Canada por las tarifas de duelo, tras la muerte de su abuela. El chatbot le dijo que podía comprar un billete a precio completo y pedir el descuento después. Cuando lo pidió, Air Canada lo rechazó, porque la política real obligaba a solicitarlo antes del viaje. Él insistió, y la aerolínea sostuvo ante el Tribunal de Resolución Civil de la Columbia Británica que no respondía de lo que hubiera dicho su propio chatbot, porque este era, según sus palabras, «una entidad legal independiente, responsable de sus propias acciones». El tribunal no lo aceptó. En febrero de 2024 condenó a Air Canada a pagar unos seiscientos cincuenta dólares canadienses por daños. La cantidad era pequeña. El argumento que la aerolínea intentó colar, no: una empresa tratando de convencer a un tribunal, en serio, de que su propio sistema de atención al cliente era algo distinto de ella misma.

Por las mismas fechas, CNET publicaba sin mucho ruido explicaciones de finanzas personales escritas en gran parte por una herramienta de IA, sin avisar claramente a los lectores. Cuando periodistas de fuera 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 forma que podría haberle costado dinero a un lector. CNET corrigió los textos y empezó a revisar el resto de su archivo asistido por IA, pero solo después de que alguien de fuera detectara los fallos. El problema no fue que la IA se equivocara de vez en cuando con las matemáticas financieras, algo que le puede pasar a cualquier

herramienta. Fue que nada dentro del proceso de CNET estaba pensado para detectarlo antes.

La cuarta instantánea no se parece a las otras tres, porque no fue un fallo. Fue una negociación. En 2023, SAG-AFTRA, el sindicato estadounidense de actores, estuvo ciento dieciocho días en huelga, en buena parte por cómo podían usar los estudios la IA para recrear la imagen y la voz de los intérpretes. El acuerdo que la cerró, ratificado en diciembre con un setenta y ocho por ciento de votos a favor, no prohibió la tecnología. Exigió consentimiento antes de que un estudio pudiera crear o reutilizar la réplica digital de un intérprete: explícito, documentado y concreto sobre el uso previsto. Exigió una compensación cuando esa réplica hacía un trabajo que de otro modo habría pagado una actuación en vivo, y protegió a los actores de fondo frente a ser sustituidos así. Y dejó escrito que el consentimiento sobrevive a la muerte del intérprete, salvo que este hubiera dicho lo contrario. Nada de esto esperó a una demanda o a un escándalo. Toda una profesión negoció sus condiciones antes de que el uso de la tecnología se volviera normal a su alrededor.

Esas cuatro instantáneas son momentos puntuales: la respuesta de un chatbot, unos párrafos escritos por IA, una cláusula de un convenio. Hay dos casos más, los dos a la escala de una empresa entera y sostenidos durante años, que muestran lo que pasa cuando los mismos términos de este libro se ponderan de forma muy distinta durante mucho tiempo. Ninguno tiene que ver con el tipo de IA del que habla el resto del libro, y uno es incluso anterior a ella. Merece la pena fijarse en que el patrón no empieza con la IA. La IA sube lo que está en juego y acelera decisiones que las empresas siempre han tenido que tomar.

Boeing tenía, se mire como se mire, una capacidad extraordinaria: talento de ingeniería, conocimiento acumulado, décadas construyendo aviones como pocas empresas del planeta. Lo que

hizo con esa capacidad en el 737 MAX es el punto en el que los términos de este libro dejan de ser abstractos.

Se añadió un sistema automático de control de vuelo, el MCAS, para corregir el comportamiento del avión en ciertas situaciones de ángulo de ataque elevado, provocadas por los cambios aerodinámicos del nuevo modelo. Se rechazó una propuesta de comprobación cruzada de seguridad basada en una velocidad aerodinámica sintética, que ya se usaba en el 787, por su coste y porque podía obligar a exigir entrenamiento en simulador, algo que Boeing estaba decidida a evitar. Boeing había convertido la ausencia de ese entrenamiento en uno de sus principales argumentos de venta, y se había comprometido a pagar a Southwest al menos un millón de dólares por avión si los reguladores lo acababan exigiendo. Un ingeniero resumió después sin rodeos la filosofía de la dirección del programa: si algo no era imprescindible para que el avión funcionara o se certificara, no se instalaba.

Dos de los pilotos técnicos del 737 MAX de Boeing ocultaron información relevante sobre el MCAS al grupo de evaluación de aeronaves de la FAA. El resultado fue que los manuales de las aerolíneas y los materiales de formación de los pilotos en Estados Unidos no mencionaban el sistema en absoluto.

Entonces cayeron los aviones. 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 conspirar para defraudar a los Estados Unidos y cerró el caso con un acuerdo de enjuiciamiento diferido de más de 2.500 millones de dólares, entre multa penal, compensación a las aerolíneas y un fondo para las familias de las víctimas. Esa cifra no incluye la flota inmovilizada, los

pedidos perdidos ni los años que costó después recuperar la confianza.

El caso Boeing es anterior al tipo de IA para el que se pensó esta fórmula, así que no voy a meter el MCAS a la fuerza en el término IA. Lo que sí pone a prueba es el principio que hay debajo, que la capacidad se vuelve peligrosa cuando fallan a su alrededor el juicio, la responsabilidad y los controles de protección. Dentro del programa, las presiones de coste y de plazos pesaron muchísimo, y las preocupaciones de seguridad no siempre tuvieron la misma fuerza a la hora de decidir. El XQ se debilitó justo donde la presión comercial empezó a influir en decisiones en las que la seguridad tendría que haber marcado el límite. Una fricción protectora concreta, la posibilidad de un entrenamiento adicional en simulador, se trató como algo que había que evitar por su coste y sus consecuencias comerciales. Y una propuesta de dispositivo de seguridad se rechazó, en parte, porque podía haber obligado a ese entrenamiento. El resultado no fue un rendimiento pobre. Una capacidad que tendría que haber servido a las personas contribuyó a una catástrofe.

La versión de Costco no tiene un gran momento dramático. Y ahí está la gracia. Jim Sinegal, cofundador de la empresa, levantó el negocio sobre una política que él mismo explicaba con claridad. Contó una vez que la mano de obra era el gasto operativo dominante de Costco, y que unos setenta centavos de cada dólar que gastaba la empresa iban a sueldos. Era una decisión deliberada, pagar a sus empleados más de lo que pagaba buena parte del comercio minorista de su entorno.

El resultado cabe en una comparación que el propio Sinegal señalaba. Entre los empleados con más de un año en la casa, la rotación en Costco rondaba el siete por ciento. En muchos otros minoristas de la época estaba entre el sesenta y el setenta.

En los términos de este libro, esa cifra carga con muchísimo. El conocimiento acumulado, el juicio entrenado y la confianza no se van por la puerta cada pocos meses, como pasa en una cadena con mucha rotación. Esa estabilidad da al trabajo en equipo, a la confianza y al saber de la casa tiempo para acumularse, en lugar de tener que reconstruirse tras cada relevo de plantilla. El XQ está en la decisión misma, en una elección sostenida de pagar y retener a la gente mejor que la norma del sector. La S solo aparece si conectas ese modelo de empleo con lo que la gente vive en su trabajo: estabilidad, pertenencia, dignidad y la posibilidad de construir una vida alrededor de algo que dure más que el siguiente turno. Menos rotación reduce, además, una fricción organizativa fácil de subestimar: contratar una y otra vez, volver a formar, perder lo aprendido. Un coste que muchos comercios dan por inevitable. Aquí no hay un suceso que haga evidente la historia. La ventaja aparece despacio, porque se acumula año tras año.

Las dos empresas no son imágenes en un espejo, pero dejan a la vista la misma pregunta: ¿qué pasa cuando el juicio responsable cuesta dinero, tiempo o ventaja a corto plazo? Boeing enseña lo que puede ocurrir cuando la presión comercial debilita el juicio y los controles. Costco enseña algo más callado, que es cómo las decisiones sobre las personas pueden reducir la fricción destructiva y conservar la capacidad humana con los años. Los resultados son distintos, pero las preguntas vuelven siempre. Quién revisa el trabajo. Quién responde cuando está mal. Qué les pasa a las personas cuyo juicio, imagen o medio de vida toca ahora el sistema. Y qué está dispuesta a sacrificar una empresa, o a negarse a sacrificar, cuando contestar a esas preguntas tiene un coste real. Nada de esto es especulación. Ya ha pasado.

Donde el patrón se rompe

Si cada fracaso de este capítulo fuera un fallo de responsabilidad y cada éxito un éxito de sentido, sería razonable pensar que XQ y S son la misma variable con dos disfraces. No lo son. Y un marco que siempre sale bien parado vale poco. Por eso aquí van tres casos que van a contracorriente de todo lo anterior.

El primero es un sistema que hizo casi todo lo que pide la responsabilidad. Durante las dos últimas décadas, la medicina estadounidense pasó sus historias clínicas al ordenador, empujada por incentivos federales y por buenas razones: órdenes legibles, registros de auditoría, menos expedientes perdidos, constancia de quién decidió qué y cuándo.

Una historia clínica electrónica es el expediente del paciente en formato digital: diagnósticos, pruebas, tratamientos, notas del médico. Cada entrada queda registrada con su autor y su hora. Pocos sistemas son tan trazables.

En 2016, un equipo dirigido por Christine Sinsky, de la Asociación Médica Estadounidense, observó durante 430 horas a 57 médicos de cuatro especialidades. En la jornada de consulta, los médicos pasaron el 27 por ciento de su tiempo cara a cara con los pacientes y el 49 por ciento con la historia electrónica y el trabajo de despacho. Por cada hora con un paciente, casi dos más delante de la pantalla. Los que llevaron un diario apuntaron además entre una y dos horas cada noche, casi siempre con esos mismos registros. Una encuesta nacional dirigida por Tait Shanafelt, de la Clínica Mayo, publicada ese mismo año, encontró que los médicos que usaban historias electrónicas y sistemas informatizados de órdenes médicas estaban menos satisfechos con el tiempo que les robaba el papeleo, y que usar esos sistemas de órdenes se asociaba a un mayor riesgo de desgaste profesional.

Nadie actuó de mala fe en esta historia. Pero hay que precisar qué era real. Tener las decisiones registradas no prueba por sí solo una responsabilidad moral alta. Prueba trazabilidad, y la trazabilidad puede convivir con el agotamiento. XQ pregunta hacia dónde apunta un sistema y quién responde de él, y un registro de auditoría perfecto solo contesta a la segunda mitad. Lo que se fue perdiendo fue la parte del trabajo que lo hacía vocación: sentarse con una persona y escucharla. Trazabilidad alta, S en caída. Es exactamente el caso para el que se construyó esta fórmula, y un marco con un solo término moral lo pasaría por alto.

El segundo caso es una fricción protectora que se mantuvo y aun así falló. Los monitores de los hospitales son una salvaguarda de manual, una máquina que no se cansa nunca, vigilando el momento en que el corazón o la respiración de un paciente empiezan a fallar.

En abril de 2013, la Joint Commission, el organismo que acredita a los hospitales de Estados Unidos, lanzó una alerta sobre la seguridad de las alarmas. Su base de datos recogía 98 incidentes relacionados con alarmas entre enero de 2009 y junio de 2012. En 80 de ellos, el paciente murió. El factor que más se repetía era la fatiga de alarmas.

La fatiga de alarmas es lo que pasa cuando un aviso suena tantas veces que la gente deja de oírlo. Los estudios citados en la alerta calculaban que entre el 85 y el 99 por ciento de las señales de alarma no requieren ninguna intervención clínica. El personal aprende a ignorarlas, a bajarles el volumen o a apagarlas, y la alarma que de verdad importa suena igual que todas las demás. Como la notificación era voluntaria, la cifra real es casi con seguridad mayor.

La salvaguarda estaba ahí. Nadie la quitó. Se multiplicó hasta que dejó de significar nada. Es la decisión del 27 de junio de la Parte III, a la escala de todo un sector. La fricción protectora puede fallar porque se pierde. También puede fallar porque se ahoga.

El tercer caso va en sentido contrario, porque es un control que se eliminó pese a objeciones serias, y el daño temido no llegó. Durante décadas, los grandes aviones comerciales volaron con tres personas en cabina. La tercera era el ingeniero de vuelo, encargado de vigilar y gestionar los sistemas del avión: motores, combustible, electricidad, presurización.

Cuando los fabricantes diseñaron una nueva generación de aviones para dos pilotos, el sindicato se opuso por considerarlo inseguro, y en 1981 se nombró un grupo de trabajo presidencial para decidirlo. Su informe, entregado el 2 de julio de 1981, concluyó que los nuevos aviones podían volar con seguridad con dos tripulantes. Buena parte del trabajo del tercer asiento ya se había trasladado a motores mucho más fiables y a sistemas automáticos. La cabina de dos pilotos se convirtió en el estándar del sector y lo sigue siendo.

La lección no es que quitar controles sea seguro. Es más estrecha y más útil, y dice que un control se puede quitar con seguridad cuando la excepción para la que existía se ha ido de verdad a otro sitio, y no antes. Antes de eliminar un paso, pregunta adónde ha ido su excepción. Si no sabes contestar, todavía no has terminado el análisis.

Pon estos tres casos junto a los demás y el marco deja de ser un paquete cerrado. La responsabilidad y el sentido son decisiones distintas. La fricción protectora puede fallar aunque la conserves. Quitar fricción puede ser lo correcto. Son afirmaciones que cualquiera puede poner a prueba y discutir, y eso es justo lo que quiero que sean.

Tres decisiones que me tocaron a mí

Antes de pasarte la silla, te debo algo. A lo largo de este libro he hecho preguntas incómodas a empresas que no conozco por dentro. Es justo que cuente tres veces en que me tocó hacérmelas a mí. Sin nombres, porque las personas de estas historias no eligieron salir en un libro.

La primera empezó con un favor. Un alto directivo me abrió la puerta para entrar en una empresa. Después quiso que siguiera sus normas. Yo tenía un encargo concreto y un equipo, y mi responsabilidad era con ellos y con lo que se me había encargado. Me centré en eso. El conflicto con aquella persona se quedó ahí.

Le debía la entrada. No le debía el criterio. Es la pregunta de XQ en su forma más pequeña: ¿a quién respondes? A veces responder a quien toca deja un conflicto abierto con quien te ayudó, y hay que aceptarlo.

La segunda ocurrió cuando era responsable de implementación y servicios de un país. Un partner y un técnico quisieron llevarse los servicios a una empresa suya, en lugar de dejarlos donde debían estar, en los servicios del fabricante. Me pidieron que pasara todos los servicios por esa empresa.

Me negué.

Aquí las preguntas de la Parte V casi se contestan solas. ¿Quién se llevaba el beneficio? Ellos. ¿Dónde tenían que estar esos servicios? Con el fabricante. Un XQ bajo casi nunca llega con ese nombre. Suele llegar como una petición.

La tercera es la que menos me gusta contar. Tuve que despedir a un empleado porque había dejado de hacer su trabajo. Acababa de divorciarse y estaba pasando por una situación personal difícil. Aun

así, había dejado de cumplir con sus obligaciones, y alguien tenía que decidir.

Decidí yo. Fue desagradable.

Es el caso que ninguna fórmula resuelve, y prefiero decirlo. La dignidad que le debía a él, por lo que estaba viviendo, y lo que le debía al resto del equipo tiraban en direcciones opuestas. Ninguna fórmula me habría dado la respuesta. Lo que sí hace esta es obligar a mirar las dos cosas a la vez, sin fingir que una de ellas no existe.

Tres decisiones, tres preguntas distintas: a quién respondes, quién se beneficia, qué le debes a la persona que tienes delante. Ahora te toca a ti.

Una decisión que te toca

Lo que sigue no ocurrió. Es una escena inventada, y la presento así a propósito, porque en ella quien decide eres tú.

Diriges el servicio de atención al cliente de una aseguradora mediana. Doscientas personas atienden llamadas y correos. Muchas llevan quince años en la casa y se saben de memoria los casos raros: la viuda que no sabe qué póliza tiene, la reclamación que no encaja en ningún formulario. Te presentan un sistema de IA capaz de resolver por sí solo cuatro de cada cinco consultas. Las esperas bajarían de veinte minutos a menos de uno. El ahorro, varios millones al año. Las pruebas han ido bien y los clientes que lo han probado están contentos.

Sabes otras cosas. Que la quinta consulta, la que el sistema no resuelve, suele ser la más difícil y la más humana. Que si el equipo pasa de doscientas personas a cuarenta, la experiencia de las otras ciento sesenta se irá con ellas. Y que dentro de tres años nadie

recordará cómo se resolvían a mano los casos que hoy solo ellas saben resolver.

Todo lo que he contado en este libro está encima de esa mesa: capacidad real, beneficios reales, una fricción que parece desperdicio y a lo mejor no lo es, un sentido que puede mudarse o evaporarse. No voy a decidir por ti. Te dejo las preguntas que me haría yo.

> *Si la decisión fuera tuya, ¿qué necesitarías saber antes de aprobarla? ¿Qué parte del ahorro dedicarías a quienes tendrán que cambiar de trabajo? ¿Y quién atendería el caso que el sistema no sabe resolver?*

Si has contestado a alguna de las tres, ya has usado la fórmula. La primera es una pregunta de XQ. La segunda, de S. La tercera, de fricción protectora. Y lo que el próximo capítulo tiene que preguntarse es qué pasa cuando estos patrones dejan de limitarse a un sector o a una empresa y empiezan a cambiar para qué sirve una economía.

> *La responsabilidad, el sentido y la fricción son decisiones distintas, y cada una puede fallar por su cuenta.*

PARTE IX

Lo que viene después del crecimiento

Una pregunta que las máquinas no pueden responder por nosotros

> *«No es poco el tiempo que tenemos, sino mucho el que hemos perdido.»*
> *Séneca, Sobre la brevedad de la vida*

Al principio te pedí que llevaras contigo una pregunta: ¿qué organiza un martes por la mañana cuando el trabajo ya no lo hace? Aquí deja de ser retórica.

Cada parte del libro ha seguido alguna versión del patrón de la Parte I, en el que una tecnología multiplica lo que una persona puede producir, y la sociedad se pasa una generación, a veces varias, averiguando cómo repartir lo creado sin romperse por el camino. Quiero terminar mirando ese patrón a la escala más grande que puedo imaginar con honradez, a treinta o cuarenta años vista. No sé si algo de esto pasará tal como lo cuento. Pero versiones de este patrón se repiten desde que la agricultura creó el primer excedente estable, y creo que también aquí merece tomarse 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 aprietan con fuerza el primer eslabón de esa cadena.

Imagina el año 2060. No hace falta imaginar un mundo en el que nadie trabaje. Basta con uno en el que la IA, la robótica y la automatización nos permitan producir dos o tres veces más con muchas menos horas de trabajo humano que ahora. Eso podría crear un problema estructural, el de una capacidad de producción enorme junto a unos ingresos salariales demasiado escasos en los hogares corrientes para absorber lo que la economía es capaz de 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 parte cada vez mayor de la producción sale del capital (las máquinas, los modelos, la infraestructura) y una parte cada vez menor de los ingresos sale del trabajo humano, el poder de compra tiene que llegar a la gente por otra vía. Si no, el sistema se vuelve cada vez más inestable, y no solo en términos morales, también en los suyos propios.

Eso no significa necesariamente el fin del capitalismo. A mí me resulta más fácil imaginar un híbrido. Empresas privadas, inversión, propiedad y competencia seguirían existiendo, junto a mecanismos mucho más fuertes para repartir lo que producen el capital, la automatización y la IA, como una renta garantizada, jornadas laborales más cortas, servicios públicos universales, fondos de inversión soberanos o públicos, impuestos sobre los rendimientos extraordinarios del capital tecnológico, más propiedad en manos de los trabajadores o alguna forma de dividendo social. Lo más probable es que fueran varias soluciones superpuestas, no una sola. La vieja etiqueta de capitalismo contra socialismo puede decirnos menos que una pregunta más concreta: quién es dueño de los sistemas que producen y quién recibe los ingresos que generan.

Desde aquí veo dos futuros muy distintos, los dos posibles. Uno es una especie de hipercapitalismo. La propiedad de la IA se concentra

de golpe, y un número pequeño de empresas, fondos y personas controla los modelos, la robótica, la infraestructura energética, los datos y casi todo el capital productivo, mientras el poder de negociación del trabajo humano sigue encogiendo. La productividad podría dispararse. Sin cambios en la propiedad o en el reparto, la desigualdad se dispararía con ella. En esa versión, el progreso se agria y se convierte en concentración.

El otro se parece más a un capitalismo social. La propiedad privada sigue, pero la sociedad acepta que, cuando el trabajo humano deja de ser la vía principal por la que los hogares reciben ingresos, el acceso a esos ingresos no puede depender solo de tener un empleo. Con un arreglo así, a nadie 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, infraestructuras y ciencia que ninguna empresa creó sola. Es un cambio más profundo de lo que suena. Y no es pura hipótesis. Alaska paga desde 1982 a cada residente un dividendo anual con cargo a su Fondo Permanente, financiado con los ingresos del petróleo. Suele estar entre 1.000 y 2.000 dólares al año. Es poco comparado con un salario digno, pero es un precedente de décadas, casi sin polémica, de un ingreso ligado a la propiedad compartida de un recurso y no a un empleo.

Carlota Pérez, cuya investigación ya apareció en la Parte I, descubrió que las revoluciones tecnológicas anteriores tardaban entre cuarenta y sesenta años en extenderse por la economía, pasando por la instalación, la crisis o punto de inflexión y el despliegue, mientras las instituciones se iban adaptando. Si la IA marca una nueva aceleración que empezó hacia 2020, sus consecuencias institucionales podrían seguir desplegándose hasta bien entrada la segunda mitad del siglo. No pretendo conocer el calendario. La pregunta política podría ir cambiando por el camino. Del cómo

crear empleos suficientes al porqué: por qué alguien necesita un empleo convencional para participar en la economía de una sociedad capaz de producir abundancia con cada vez menos trabajo humano.

Aun así, no creo que el problema más difícil sea económico. Es la S otra vez, la del sentido. El trabajo nunca ha servido solo para repartir ingresos. Durante generaciones también ha repartido identidad, relaciones, una estructura para las horas del día y un motivo para levantarse una mañana cualquiera. Hasta la jubilación saca a menudo parte de su sentido de llegar después de décadas ordenadas por el trabajo. El lado material tiene al menos un mecanismo que puedo describir a grandes rasgos: la productividad, canalizada a través de la propiedad o de los impuestos, hacia alguna forma de dividendo social y, de ahí, hacia el poder de compra. La pregunta difícil llega justo detrás, y ningún mecanismo la contesta solo: si el trabajo deja de organizar nuestra existencia, ¿qué organiza un martes por la mañana?

No tengo una respuesta general, y desconfío de quien diga tenerla. Pero sé cómo es una respuesta real, porque formo parte de ella.

En Alcoy, la ciudad de las montañas de Alicante donde tengo raíces profundas, la Fiesta de Moros y Cristianos la sostienen 28 filaes. Una filà es una compañía festera, un grupo de personas que se preparan, ensayan, se visten y se reúnen para los días de abril en honor a Sant Jordi, y durante buena parte del año que lleva hasta ellos. La fiesta es de Interés Turístico Internacional desde 1980, y en 2026 la ciudad celebró los 750 años del patronazgo de Sant Jordi. Nada de eso le paga el alquiler a nadie. Y a miles de personas les organiza muchísimos martes.

Un detalle lo explica mejor que cualquier teoría. Los actos principales de la fiesta tardan entre uno y cuatro años en prepararse. Años de reuniones y de ensayos para unas horas en la calle. Quien

los prepara no cobra por ello. Y es un trabajo que ninguna máquina podría simplificar, porque lo que se prepara no es solo un desfile. Es el tiempo que esas personas pasan juntas.

Eso no es una política. No se puede extender por decreto y no arregla el problema de los ingresos, que es otro. Pero demuestra que lo que describe la S existe fuera del trabajo pagado: un lugar al que la gente va, un oficio que practica, unas personas que notan cuando falta alguien. La pregunta es si construiremos, protegeremos y financiaremos más lugares así antes de necesitarlos, o si descubriremos lo que valían cuando el trabajo que ordenaba nuestros días ya se haya ido.

Es el mismo hilo que recorre todo el libro. Una tecnología crea un excedente y, al principio, la propiedad tiende a concentrarlo. La tensión se va acumulando hasta que, con el tiempo, las instituciones responden. Versiones de esa secuencia aparecieron tras la agricultura, durante la industrialización y en las olas tecnológicas posteriores. Las olas anteriores acabaron creando grandes categorías nuevas de trabajo humano, aunque la transición fuera dolorosa y desigual, y para la mayoría de la gente el salario siguió siendo la vía principal por la que el crecimiento llegaba a casa. Con la IA y la robótica avanzada existe, esta vez, al menos una posibilidad razonable de que el salario por sí solo ya no baste.

En esa historia hay una distinción escondida que quiero nombrar, porque el resto del capítulo depende de ella. Cuando una tecnología se lleva un trabajo, el sentido que iba unido a él puede seguir dos caminos. Puede migrar, si aparecen trabajos nuevos y la identidad, la destreza y la pertenencia se mudan a ellos. O puede evaporarse, si el trabajo desaparece y nada ocupa su lugar en la vida de quien lo hacía. Lo llamo migración del sentido y evaporación del sentido. Lo incómodo es que las dos cosas pueden pasar a la vez. Con las generaciones, el sentido migró después del telar mecánico. Para el

tejedor manual, se evaporó. Los optimistas tienen razón en que la economía siempre ha encontrado trabajo nuevo, y las páginas que siguen le dan a ese argumento todo su peso. La pregunta que plantea la S es si lo encuentra para las mismas personas, en la misma vida.

La objeción evidente merece decirse de frente. Cada ola de automatización iba supuestamente a acabar con el trabajo disponible, y cada vez, en una o dos generaciones, aparecieron categorías enteras de empleo que antes no existían y que nadie habría sabido nombrar de antemano. El economista David Autor ha hecho la defensa más completa de esa idea, y sostiene que, históricamente, la automatización ha destruido tareas concretas, no el empleo en conjunto, y la productividad que libera ha creado sistemáticamente más empleo del que ha eliminado («Why Are There Still So Many Jobs?», 2015). Su trabajo más reciente va más allá. Sostiene que la IA, en concreto, podría ayudar a reconstruir el empleo de cualificación media que las olas anteriores vaciaron, en lugar de borrar todavía más («Applying AI to Rebuild Middle Class Jobs», 2024). Él mismo advierte que ese resultado depende de que la IA se oriente deliberadamente a potenciar la pericia escasa y no a sustituirla. Es una posibilidad real, no una garantía que venga de serie con la tecnología. Los economistas tienen incluso un nombre para quien apuesta contra ese patrón por completo: la falacia de la cantidad fija de trabajo.

La renta básica universal tiene sus propios problemas, además de su promesa. Financiarla a un nivel que sustituya los salarios perdidos es caro a la escala de una economía entera, y puede debilitar las ganas de trabajar justo donde el trabajo todavía importa. Sus resultados reales, además, son más modestos de lo que dicen los discursos de uno y otro bando. Finlandia hizo lo más parecido a un experimento controlado a escala nacional: 2.000 personas en paro recibieron 560 euros al mes sin condiciones durante dos años. El informe final

encontró un efecto pequeño sobre el empleo, unos seis días más de trabajo de media al año, junto a una mejora real del bienestar declarado y de la seguridad económica percibida.

Me tomo todo eso en serio, y nada de ello es la razón por la que apostaría a que esta vez es distinto. Lo que cambia es el abanico de tareas que un solo sistema puede absorber hoy, en todos los sectores a la vez, sin los años de reconversión y de infraestructura que las olas anteriores necesitaron para llevar el trabajo desplazado hacia una categoría nueva de empleo. Eso no demuestra que el patrón histórico no vaya a repetirse. Es un motivo para no dar por hecho que se repetirá solo, 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 sea si el capitalismo desaparece. Creo que es otra, la de qué le pasa al capitalismo cuando el trabajo humano deja de ser la vía principal por la que se reparte la riqueza que genera. Si llega ese momento, dudo que traiga el fin de los mercados o de la propiedad privada. Puedo imaginar mercados y propiedad privada conviviendo con un reparto mucho más social que el de hoy, y con la obligación de organizar fuera del trabajo pagado mucho más sentido humano del que las sociedades industriales se han acostumbrado a organizar.

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 ni siquiera sea económica. Puede que sea, simplemente, qué haremos con el tiempo que estos sistemas podrían devolvernos, y quién lo recibirá. Séneca escribió hace casi dos mil años que no nos falta tiempo, que lo perdemos. Nunca tuvimos tanto a punto de sobrarnos.

El problema nunca fue que llegara el futuro. Siempre llega. La pregunta ahora es si sabremos 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, preguntando si dos personas, o veinte mil, funcionan de verdad juntas, y cuánto les cuesta la fricción por el camino. Y termina con todo un sistema económico, con la duda de si el excedente que genera, cada vez más, un capital tecnológico concentrado se reparte con amplitud suficiente para mantener unida a una sociedad. Todo lo que hay en medio volvía una y otra vez a la misma familia de preguntas, planteadas en escenarios distintos: un laboratorio, una sala de tribunal, un hospital, una aerolínea, una fábrica de aviones, un club de compras, una civilización. De un capítulo a otro no cambia la pregunta. Cambia lo que está en juego.

La conclusión a la que sigo llegando es que la capacidad nunca fue, por sí sola, la variable que decidió el resultado. Cada historia del libro juntó una capacidad real con una mezcla distinta de los otros términos a los que no he dejado de volver: XQ, la fricción protectora, el sentido, la amplitud con que se repartió el beneficio. Y la diferencia dependió menos de cuánta capacidad había que de cómo se gestionó la responsabilidad a su alrededor. Los resultados salieron de cómo se dirigió esa capacidad, de cómo se comprobó, de cómo se compartió y del propósito digno al que se unió. De lo potente que era la herramienta, casi nunca.

Antes, buena parte de la rendición de cuentas nacía sola de las relaciones humanas que rodeaban una decisión. Un padre o una madre, un mentor, el tendero de la esquina: quien te enseñaba algo,

te vendía algo o decidía algo en tu nombre seguía estando allí después, en el mismo mundo en el que tú tenías que vivir. Eso ya no es automático. Cuando quien enseña, aconseja o decide llega de fuera del linaje humano, lo que en la Parte VI llamé transmisión extralineal, la rendición de cuentas hay que construirla a propósito, casi siempre por personas que nunca llegan a conocer a quien está al otro lado de la decisión. En los ejemplos que salieron bien, esa responsabilidad estuvo presente lo bastante pronto como para dar forma al resultado. En los fracasos, fue débil, llegó tarde o nadie le hizo caso hasta que las consecuencias ya estaban a la vista.

No creo que la tecnología sea la única parte difícil. La brecha más dura puede ser todo lo que la rodea. Las instituciones ya han absorbido antes tecnologías disruptivas, aunque despacio, de forma desigual, casi siempre después de que un daño real forzara la cuestión. Lo distinto esta vez no es que la tecnología no tenga precedentes. Es que la ventana para adaptarse parece encogerse mientras las consecuencias se extienden cada vez más deprisa, antes de que la capacidad vuelva a moverse y nos cambie el problema bajo los pies.

Así que la conclusión no es una predicción, y nunca quiso ser una advertencia. Se parece más a una instrucción que me doy a mí mismo tanto como a quien me lea. Lo más probable es que la capacidad siga avanzando más deprisa de lo que nuestro juicio y nuestras instituciones se adaptan a ella. Esa brecha es el verdadero tema de este libro. Cerrarla, aunque sea en parte, puede ser lo más útil que todavía está en nuestra mano.

Por dónde empezar

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

Si diriges un equipo, trata la fricción protectora como una decisión de diseño y 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 medio. Y cuenta tus alarmas, porque un control que ya nadie oye no protege a nadie.

Si diriges o asesoras una empresa, publica qué pueden hacer tus sistemas y qué los detiene, y revisa las dos cosas cada cierto tiempo, sin esperar a que algo salga mal. La distancia entre capacidad y gobernanza crece sin hacer ruido hasta que un incidente la deja a la vista.

Si das forma a políticas públicas o trabajas en una institución de alcance público, financia el trabajo lento y poco lucido (evaluación, auditoría, investigación en interpretabilidad, apoyo a quienes tienen que cambiar de empleo) a un ritmo más parecido al de la capacidad que avanza que al que han llevado históricamente las instituciones.

Si nada de esto te describe, el mismo instinto funciona a menor escala. Fíjate en dónde se está comprando velocidad a costa de quitar un control que nadie ha revisado en años, en tu trabajo o en tu propia casa. Es el mismo patrón de todo el libro, del tamaño que tenga cuando te lo encuentres.

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

El martes

Termino donde empecé, con la pregunta que te pedí que llevaras contigo: ¿qué organiza un martes por la mañana?

En algún lugar de Alcoy, esta misma semana, alguien está preparando algo que no saldrá a la calle hasta dentro de dos o tres años. Nadie le paga por ello. Si deja de aparecer, los demás lo notan. No sé si esa es la respuesta para todos, pero sé que es una respuesta real.

> *Hemos ganado capacidad. Todavía tenemos que decidir qué vida queremos sostener con ella.*

Glosario

Agente / sistema agéntico: Un sistema de IA metido en un bucle con herramientas, memoria y permisos: da un paso, observa el resultado, decide el siguiente y sigue así hasta que algo lo detiene, sea un límite de tiempo, una persona o una regla de parada. Partes III y IV.

Alucinación: Que un modelo de lenguaje afirme algo falso con la misma soltura con la que afirma algo cierto. Tiene dos orígenes principales: hay hechos demasiado raros o arbitrarios para aprenderse bien durante el entrenamiento, y hay errores que persisten porque el entrenamiento premia una respuesta segura por encima de un sincero «no lo sé». Parte IV.

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

Capitalismo social: El segundo futuro que imagina la Parte IX: la propiedad privada y los mercados siguen, pero la sociedad acepta que el acceso a los ingresos no puede depender solo de tener un empleo convencional, y construye mecanismos (una renta garantizada, la propiedad de los trabajadores, un dividendo social) para repartir el excedente con más amplitud.

Evaporación del sentido: Cuando el trabajo desaparece y la identidad, la destreza y la pertenencia que iban con él no encuentran otro lugar en la vida de quien lo hacía. Parte IX.

f (fricción): El denominador de la Salud del TMC y de la fórmula del Impacto: todo lo que drena en silencio lo que un equipo o un sistema podría conseguir. En la fórmula del Impacto, f es la fricción que desperdicia capacidad; los controles que evitan un daño no cuentan como desperdicio por el mero hecho de consumir tiempo, aunque también hay que revisarlos. Véanse fricción de desperdicio y fricción protectora. Partes II, III y V.

Falacia de la cantidad fija de trabajo: El nombre que dan los economistas a la idea de que existe una cantidad fija de trabajo, de modo que lo que elimina la automatización desaparece para siempre. Se ha demostrado falsa tantas veces que apostar por ella es, en sí mismo, un riesgo. Parte IX.

Fatiga de alarmas: Lo que ocurre cuando una salvaguarda suena tan a menudo que la gente deja de oírla. Según la Joint Commission, fue el factor más frecuente en las muertes hospitalarias relacionadas con alarmas. Es fricción protectora que falla porque se ahoga, no porque se elimine. Parte VIII.

Fórmula del Impacto: Impacto = [(IH + IA) × XQ × S] / f. La inteligencia humana y la artificial se suman; XQ y S multiplican el resultado; la fricción lo divide. Procede del ensayo *El hilo del crecimiento*. Es una lente para pensar, no una ecuación para calcular. Parte V.

Fricción de desperdicio: La fricción que no sirve para nada real: la reunión que existe porque una vez existió, la cadena de aprobaciones que nadie recuerda haber creado. Parte III.

Fricción protectora: La 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 lo haga más lento. Parte III. Véase también el MCAS, en la Parte VIII, para lo que ocurre cuando esa fricción se elimina por presión comercial en lugar de rediseñarse. Puede fallar aunque se conserve (véase fatiga de alarmas), y a veces un control puede eliminarse con seguridad cuando la excepción que protegía se ha trasladado a otro sitio (la cabina de dos pilotos, Parte VIII).

Hipercapitalismo: El primero de los dos futuros que imagina la Parte IX para una economía movida por la IA: la propiedad de la IA se concentra de golpe en pocas manos mientras el poder de negociación del trabajo humano sigue encogiendo.

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

Interpretabilidad: El campo de investigación que intenta entender qué pasa realmente dentro de un modelo de IA mientras produce una respuesta. Útil de verdad, y todavía parcial, según reconocen los propios investigadores. Parte IV.

MCAS (Sistema de Aumento de las Características de Maniobra): El sistema automático de control de vuelo de Boeing en el 737 MAX, implicado en dos accidentes mortales. No es IA en el sentido en que el resto del libro usa la palabra: es un tipo de automatización más antiguo, que aparece aquí porque el mismo patrón de fondo ya funcionaba mucho antes de que existiera la IA. Parte VIII.

Migración del sentido: Cuando el sentido unido a un trabajo perdido se muda a trabajos nuevos. Puede darse para una sociedad a lo largo de generaciones mientras se evapora para una persona en una sola vida. Parte IX.

Moral Machine: Experimento publicado en *Nature* en 2018 que recogió unos cuarenta millones de decisiones morales de personas de 233 países y territorios sobre dilemas de coches autónomos. Encontró preferencias compartidas en todas las culturas y diferencias claras entre tres grandes bloques: occidental, oriental y meridional. Parte V.

Neurastenia: El nombre que dio en 1869 el neurólogo George Miller Beard a un agotamiento nervioso que atribuyó a la vida moderna: el vapor, la prensa periódica, el telégrafo. El diagnóstico vivió su auge a principios del siglo XX. Parte I.

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

Preentrenamiento: La primera etapa del entrenamiento de un modelo de lenguaje: exponerlo a cantidades enormes de texto para que aprenda a predecir lo que viene a continuación. Le da buena parte de su capacidad general, no su comportamiento como asistente. Parte IV.

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

S (Sentido): En la fórmula del Impacto, la capacidad de sostener el propósito, la pertenencia y la conexión humana cuando la producción económica necesita menos tiempo de una persona. Incluye la identidad cultural de una comunidad, su idiosincrasia: su manera propia de estar juntos y de decidir qué merece la pena. Por eso es también el lugar de la fórmula donde vive la parte de la moral que cambia de una cultura a otra. Parte V; vuelve a ser central en la Parte IX. Se aplica como una lente, no se calcula como un número.

Salud del TMC: La revisión de 2025 del TMC: Salud del TMC = ((T × C × M) / f) × 100, que añade la fricción como denominador. En este libro se usa como lente, igual que la fórmula del Impacto. Parte II.

Tecnoestrés: Término que acuñó el psiquiatra Craig Brod en 1984, cuando el ordenador llegaba a las oficinas, para el estrés de no saber convivir de forma sana con las nuevas tecnologías. Parte I.

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 fracasar igual si quienes lo forman no confían, no se entienden o no se apoyan. Parte II.

Token: Una palabra con dos significados en este libro. En seguridad (Parte III), una credencial digital temporal que acredita quién pide algo y qué se le permite hacer; renovarla es refrescarla antes de que caduque. En los modelos de lenguaje (Parte IV), la unidad mínima de texto que un modelo lee y escribe: una palabra o un trozo de palabra.

Transformer: La arquitectura de la mayoría de los modelos de lenguaje punteros de hoy, presentada en 2017 en el artículo «Attention Is All You Need». Su mecanismo clave es la atención. Parte IV.

Transmisión extralineal: Término propio de este libro para un cuarto canal de transmisión cultural, junto a la vertical, la horizontal y la oblicua: la enseñanza que llega por completo de fuera del linaje humano, sin envejecimiento, sin descendencia y sin consecuencias vividas propias. Parte VI.

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

Transmisión oblicua: Transmisión cultural que llega de un adulto que no es el padre ni 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 que se absorben antes de poder juzgarlas. Uno de los tres canales clásicos (Cavalli-Sforza y Feldman, 1981). Parte VI.

XQ (Inteligencia Moral): En la fórmula del Impacto, el término que pregunta si la capacidad combinada de humanos e IA se dirige con responsabilidad, hacia resultados seguros, justos, responsables y ampliamente beneficiosos, en lugar de limitarse a escalar más deprisa. No dicta un contenido universal del bien: plantea preguntas que cualquier cultura puede contestar a su manera y ninguna debería saltarse (quién responde, quién paga, quién se beneficia, si se decidió a sabiendas, si lo sabían los afectados). Parte V. Como S, se aplica como una lente, no se calcula como un número.

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.
- George M. Beard, *American Nervousness: Its Causes and Consequences* (Putnam, 1881).
- Craig Brod, *Technostress: The Human Cost of the Computer Revolution* (Addison-Wesley, 1984).
- American Psychological Association, «2023 Work in America Survey: Artificial Intelligence, Monitoring Technology, and Psychological Well-Being», septiembre de 2023, apa.org. Encuesta realizada por The Harris Poll a 2.515 trabajadores de Estados Unidos entre el 17 y el 27 de abril de 2023.

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, acompañado del «OpenAI – Hugging Face Incident Technical Report». Conclusiones principales validadas por asesores externos contratados por OpenAI, entre ellos la firma de seguridad CrowdStrike.
- METR y Redwood Research, «Brief independent investigation of agents' behavior, reasoning and collaboration in the OpenAI / Hugging Face hacking incident», 26 de agosto de 2026, metr.org. Fuente de la cifra de unos 700 agentes participantes en el ataque.

- Emily Forlini, «OpenAI, independent firms publish reports into rogue AI agent attack on Hugging Face. Here's what they say—and what they don't», *Fortune*, 26 de agosto de 2026. Sobre la presentación de OpenAI en Black Hat del 5 de agosto de 2026.
- «OpenAI saw warning signs weeks before Hugging Face breach», *Axios*, 26 de agosto de 2026. Sobre el reconocimiento, en el informe técnico, de las señales tempranas.
- 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.
- «OpenAI pauses some training amid allegations its rogue agents behaved more badly than first thought», *The Register*, 28 de septiembre de 2026. Sobre la segunda pausa del entrenamiento.
- «OpenAI is pausing training of its most powerful AI models after rogue agents hit government sites», *Quartz*, 28 de septiembre de 2026. Sobre los accesos a webs de organismos públicos de Estados Unidos y de Australia.
- «OpenAI abandons plan to release upcoming model as safety concerns escalate», *CNBC*, 28 de septiembre de 2026. Sobre GPT-6.1 Astra; la noticia la adelantó *The Wall Street Journal*.
- «White House Accord on Super Intelligence. Joint Commitment on Frontier Responsibilities», 29 de septiembre de 2026, publicado por Donald Trump en Truth Social. Texto íntegro en *Washington Examiner*.
- «Trump and major AI executives sign "morally binding" voluntary controls», *CBS News*, 29 de septiembre de 2026.
- «OpenAI is sued over rogue AI Hugging Face cyberattack», *CNBC*, 30 de septiembre de 2026. Sobre la demanda de Legal Advocates for Safe Science and Technology.

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 al escribir poemas, la «universalidad conceptual» entre lenguas y los circuitos de aritmética en paralelo.
- Adam Tauman Kalai y Ofir Nachum (OpenAI), «Why Language Models Hallucinate», arXiv:2509.04664, septiembre de 2025; openai.com/index/why-language-models-hallucinate.
- Conversación del autor con un asistente de inteligencia artificial de uso general, 4 de octubre de 2026. Testimonio propio; capturas de pantalla y registro de la conversación en poder del autor.

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 mundial de los centros de datos.
- Agencia Internacional de la Energía, «Key Questions on Energy and AI», 2026, iea.org/reports/key-questions-on-energy-and-ai. Consumo eléctrico de los centros de datos (485 TWh en 2025 y unos 950 TWh en 2030) e inversión de las empresas tecnológicas frente a la inversión en petróleo y gas.
- 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.
- Edmond Awad, Sohan Dsouza, Richard Kim, Jonathan Schulz, Joseph Henrich, Azim Shariff, Jean-François Bonnefon e Iyad Rahwan, «The Moral Machine experiment», *Nature* 563, n.º 7729 (2018): 59–64.

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. 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.
- Premio Nobel de Química 2024, concedido a David Baker, Demis Hassabis y John M. Jumper. Materiales de prensa de la 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» (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 con IA en CNET y sus correcciones posteriores, *Futurism*, enero de 2023.
- SAG-AFTRA, contratos de televisión y cine de 2023, disposiciones sobre inteligencia artificial.
- Departamento de Justicia de los Estados Unidos, acuerdo de enjuiciamiento diferido con 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 de denunciantes, *The Seattle Times*, 2019.

- MIT Sloan School of Management, «How Costco's obsession with culture drove success», *Ideas Made to Matter*, mitsloan.mit.edu. Jim Sinegal sobre los salarios (setenta centavos de cada dólar gastado) y la rotación.
- Christine Sinsky et al., «Allocation of Physician Time in Ambulatory Practice: A Time and Motion Study in 4 Specialties», *Annals of Internal Medicine*, septiembre de 2016.
- Tait D. Shanafelt et al., «Relationship Between Clerical Burden and Characteristics of the Electronic Environment With Physician Burnout and Professional Satisfaction», *Mayo Clinic Proceedings* 91, n.º 7 (2016): 836–848.
- The Joint Commission, «Medical device alarm safety in hospitals», *Sentinel Event Alert*, n.º 50, abril de 2013.
- Report of the President's Task Force on Aircraft Crew Complement, 2 de julio de 1981 (disponible a través del Defense Technical Information Center, ADA106889).

Parte IX

- Séneca, *De brevitate vitae* (Sobre la brevedad de la vida), 1, 3: «Non exiguum temporis habemus, sed multum perdidimus». Traducción del autor.
- 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).
- Associació de Sant Jordi, Alcoy, asjordi.org. La Fiesta de Moros y Cristianos, sus 28 filaes y su declaración de Interés Turístico Internacional (1980); 750 aniversario del patronazgo de Sant Jordi, 2026.
- 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.