

LA SERIE DIACROMA · LIBRO UNO

La computadora dice que eres tú

*Qué pasa cuando no queda nadie a quien
preguntar*

Copyright © 2026 Jose Miguel Calderón Carpio.

Todos los derechos reservados. Ninguna parte de este libro puede reproducirse ni transmitirse de ninguna forma sin autorización previa por escrito, salvo citas breves en reseñas.

Este libro se basa en hechos que fueron reportados públicamente al momento de escribirlo. Cuando se nombra a personas, el relato sigue ese reporte público; las caracterizaciones de asuntos en disputa se identifican como tales.

Verit Global Labs · Primera edición, 2026

Índice

Algo decide sobre tu vida. ¿Sabes qué?

MOVIMIENTO I · LA CAJA NEGRA

- Uno Cuando una máquina te rechaza, ¿a quién queda por preguntar?
- Dos ¿Por qué está tan segura de sí misma, y se equivoca tan fácil?
- Tres Si puede escribir un correo perfecto, ¿por qué le confiaríamos decidir quién se queda con la cama del hospital?

MOVIMIENTO II · NO SE PUEDE PARCHAR

- Cuatro ¿Por qué no simplemente lo arreglan?

MOVIMIENTO III · EL ACTO QUE NO PUEDES DESHACER

- Cinco ¿A quién queda apelar, una vez que ya ocurrió?

MOVIMIENTO IV · ¿A DÓNDE SE FUE EL HUMANO?

- Seis Antes había alguien a quien podías llegar. ¿Dónde está ahora?

MOVIMIENTO V · LA REGLA QUE IGNORÓ, Y LA CONFESIÓN QUE NO ES UNA RAZÓN

Siete Si rompió la regla que le diste, ¿de qué sirvió la regla? Y cuando se explica, ¿es una razón, o solo más palabras?

MOVIMIENTO VI · LO QUE YA SABEMOS

Ocho ¿Qué hace una persona cuidadosa, justo antes de lo que no puede deshacer?

Nueve ¿Y cómo se ve ya una decisión justa sobre una persona?

EL ARGUMENTO

Algo decide sobre tu vida. ¿Sabes qué?

En noviembre de 2019, un desarrollador de software llamado David Heinemeier Hansson publicó lo que parecía, en su momento, una queja menor.

Él y su esposa habían solicitado, cada uno, la nueva tarjeta de crédito de Apple. Declaran impuestos de forma conjunta. Comparten su dinero y su vida. Según la medida que se supone que más importa, el puntaje de crédito de ella era el más alto de los dos. Y la tarjeta le ofreció a él un límite de gasto veinte veces mayor que el de ella.

Estalló como una bengala. Miles de personas respondieron: *a nosotros nos pasó lo mismo*. El propio cofundador de Apple, Steve Wozniak, dijo que su límite había salido diez veces mayor que el de su esposa. La palabra para nombrarlo llegó casi al instante, y era la palabra obvia. *Sexista*. Un algoritmo sexista, ordenando a los hombres por encima de las mujeres. El regulador financiero de Nueva York abrió una investigación. La historia se contaba sola.

Solo que esa no es la historia que encontraron los investigadores. Dieciséis meses y unas cuatrocientas mil

cuentas después, el regulador concluyó que no había habido discriminación ilegal. Hombres y mujeres con perfiles crediticios similares habían recibido, en general, ofertas similares. Así que la versión viral era, en el sentido legal estricto, falsa.

Pero detente en lo que ese mismo informe encontró por debajo de la versión que todos habían estado gritando, porque ahí estuvo la verdadera historia todo el tiempo. Los investigadores encontraron que el banco no podía explicar adecuadamente sus propias decisiones. Cuando Hansson y su esposa llamaron para preguntar por qué él, por qué veinte veces, por qué el puntaje de ella no la había salvado, las personas al teléfono no podían decírselo. No *que no quisieran*. Que *no podían*. La respuesta que finalmente obtuvieron, según Hansson, fue que esto es simplemente así, y tienes que aceptarlo.

Puede que esa sea la frase más honesta que alguien dijo en todo el asunto. No porque quienes la dijeron fueran perezosos o crueles. Porque era verdad. Un sistema había hablado, y al sistema no se le podía hacer decir por qué. Ni al cliente, ni al agente de servicio, ni, de ninguna forma que satisficiera a una persona a la que se le debe una razón, al propio banco. Se había tomado una decisión sobre la vida de dos personas, y en algún punto entre los datos y el veredicto, la razón simplemente se había evaporado. Había un número. El número no podía hablar.

Ahora pon esa historia pequeña, casi cómica (un límite de crédito, un tuit viral, una pareja que, al final, podía costear el

inconveniente) junto a lo que ocurría en esos mismos años, en un país con uno de los gobiernos más confiables del mundo.

En los Países Bajos, la autoridad tributaria operaba un sistema para atrapar a padres que hacían trampa con los subsidios de cuidado infantil. Puntuaba a las familias por el riesgo de fraude, y entre las cosas que sopesaba estaba si un padre tenía una segunda nacionalidad. Un puntaje alto no abría una investigación que luego juzgara un ser humano. El puntaje *era* el juicio. Tus subsidios se detenían. El Estado exigía años de pagos de vuelta, a menudo decenas de miles de euros, y te entregaba una tarea imposible: probar que no habías cometido un fraude del que nadie te había dicho nunca que se te sospechaba, usando papeleo que la oficina tributaria con frecuencia ya había perdido.

Una de esas madres era una mujer llamada Chermaine Leysner. En 2012 estudiaba para trabajadora social, criando a tres niños pequeños, cuando llegó una carta exigiéndole devolver su subsidio de cuidado infantil de años atrás. No era un caso raro. Una investigación parlamentaria posterior, cuyo informe se tituló, sin suavizarlo, *Injusticia sin precedentes*, encontró que lo que desataba estas cascadas ruinosas podía ser algo tan pequeño como un solo error administrativo, una casilla, una firma faltante. Encontró que la autoridad había violado los principios básicos del Estado de derecho. Para cuando terminó, unas veintiséis mil familias habían sido señaladas erróneamente como defraudadoras. Hubo gente que quebró. Gente que perdió su casa y su empleo. Más de mil

niños fueron separados de sus padres y colocados en hogares de acogida. Algunos padres se quitaron la vida.

En enero de 2021, todo el gobierno neerlandés renunció por ello.

Mira las dos historias lado a lado, porque son la misma historia contada a dos volúmenes. Un número tratado como una razón. Nadie, ni el cliente, ni el trabajador del caso, ni quienes construyeron la cosa, que pudiera ponerse de pie y decir por qué. Y ningún registro después que la persona en el extremo equivocado pudiera usar para defenderse. La única diferencia real entre la tarjeta de crédito y la catástrofe era el tamaño del desastre. Una era un hombre con un límite de gasto que no necesitaba. La otra eran decenas de miles de familias empujadas a la ruina, sus hijos arrancados de ellas, su gobierno derribado.

. . .

En algún lugar que no puedes ver, algo está decidiendo sobre tu vida. Tal vez si tu beca sale. Tal vez si una persona real llega a leer tu solicitud de empleo, o si queda filtrada un paso antes de que alguien lo hiciera. Cada día, las organizaciones toman miles de estas pequeñas decisiones sobre personas, y una máquina ahora toma cada vez más de ellas, en menos de un segundo, donde nunca lo verás ocurrir.

Hemos aprendido a confiar en las máquinas, y a menudo con razón. Cuando hay una respuesta correcta contra la cual verificar (el tumor en la radiografía, las palabras en la grabación, la única tarjeta robada escondida entre un millón de

honestas) una máquina puede ganarnos, y podemos *probar* que nos ganó. Hay una clave contra la cual calificarla.

Una decisión sobre una persona no es ese tipo de problema. No hay una respuesta que consultar. Solo hay una elección, y la correcta depende de quién es esa persona: qué puede cargar, a qué ha accedido, qué permiten las reglas, qué le costará la próxima semana. No puedes verificarla contra una clave, porque no hay clave. La pareja de la Apple Card y Chermaine Leysner no se toparon con matemáticas rotas. Se toparon con una máquina haciendo con confianza el único trabajo para el que las máquinas nunca fueron construidas, y con un muro de personas que ya no podían explicar la máquina de la que habían llegado a depender.

Casi todo el ruido sobre la inteligencia artificial es sobre mejores respuestas, más rápidas y que se leen más como si las escribiera una persona. Creo que el problema, y la única salida, empieza en un lugar completamente distinto. Con mejores *preguntas*.

Las tomaremos despacio, una a la vez, con espacio para sentarse dentro de cada una. Para la última, habrás visto la forma del problema entero, y estarás sosteniendo la única pregunta más difícil que deja atrás. Acertar esa pregunta, que es todo lo que este libro se propone, resulta ser la mayor parte del trabajo. Las respuestas pueden venir después.

MOVIMIENTO I

La caja negra

UNO

Cuando una máquina te rechaza, ¿a quién queda por preguntar?

Hay un tipo de *no* antiguo y corriente que todos hemos sobrevivido. Un casero te rechaza, una escuela, un banco, un jefe. Duele, y a veces es injusto, pero viene con algo en lo que quizá nunca notaste que te apoyabas. Viene de una persona. Y a una persona se le puede preguntar por qué. Puedes plantarte en la puerta y exponer tu caso. Puedes señalar lo que se les pasó. Puedes encontrar al que está por encima y que sí escuchará, o la ley que dice que tienen que hacerlo. El *no* tiene un autor, y a un autor se le puede llegar.

Esta es una historia sobre lo que le pasa a ese sencillo acto humano cuando el autor desaparece.

En la primavera de 2021, una mujer llamada Mary Louis salió a buscar departamento. Era el tipo de solicitante por el que un casero debería competir. Había pagado renta durante años. Tenía un vale de vivienda, del tipo que financia el gobierno, lo que significaba que la mayor parte de la renta de cada mes llegaría a la cuenta del casero directamente de una agencia pública, a tiempo, tuviera o no Mary un mes difícil. El vale era, en un sentido real, una promesa del Estado de que ella

respondía por la renta. Encontró un lugar en Massachusetts que funcionaría. Llenó la solicitud. Luego esperó, de esa forma particular en que una persona espera cuando un techo depende del sí de un extraño.

La respuesta llegó como un número. Trescientos veinticuatro. Ese era su Puntaje SafeRent. SafeRent es una empresa que evalúa inquilinos por cuenta de los caseros, y había pasado los datos de Mary por su software y devuelto una sola cifra, en una escala que va de 200 a 800. Venía con un informe, de once páginas. En ninguna de esas once páginas se decía cómo se había llegado al 324, qué hechos habían entrado en él, cuáles habían pesado más, ni qué podría haber hecho una persona en la posición de Mary para salir más alto. El informe no explicaba. Puntuaba. Trescientos veinticuatro estaba por debajo de la línea, y el departamento se esfumó.

Así que Mary hizo lo más natural del mundo. Trató de averiguar por qué, y de pedir que alguien revisara su caso otra vez. La respuesta que recibió de la empresa que administraba el edificio merece leerse despacio, porque es el sonido exacto del mundo del que trata este libro. *No aceptamos apelaciones, decía, y no podemos anular el resultado de la Evaluación de Inquilino.*

Léelo dos veces. *No hemos revisado su caso y lo sostenemos. No esto es lo que salió mal, y esto es lo que lo cambiaría. No podemos anular el resultado.* Las personas que estaban entre Mary y el departamento le decían, con honestidad, que no eran ellas quienes habían decidido. La decisión se había tomado en algún lugar al que no podían llegar, por algo que no podían

cuestionar, y su trabajo era solo leerla en voz alta. Había un número. El número había hablado. Y nadie en esa oficina tenía el poder de hacerlo hablar de nuevo.

Ahora sostén la parte que debería quitarte el sueño. Mary no había hecho nada malo. El vale en su mano garantizaba que una agencia pública pagaría la gran mayoría de la renta, directo al casero, cada mes, a tiempo. Ese es casi el hecho más importante sobre si Mary puede mantener un techo sobre su cabeza, y el puntaje no lo pesó en absoluto. Se apoyó, en cambio, en su historial de crédito, una cifra moldeada a lo largo de una vida de tener menos, y convirtió eso en un veredicto sobre si merecía un hogar. Lo único construido para protegerla era invisible para la máquina. Lo que contó en su contra fue el residuo de ser pobre.

“Todo se basa en números”, dijo Mary después. “No recibes de ellos la empatía individual.” Y luego, más bajo, con la voz de quien ha entendido exactamente lo que le hicieron: “No hay forma de vencer al sistema. El sistema siempre nos va a vencer.”

Cuando el caso llegó por fin a un tribunal, la propia defensa de SafeRent delató todo el juego. La empresa argumentó que no se le debía culpar, porque no había tomado la decisión final. Eso lo hizo el casero. Pero voltea hacia el casero, y el casero señala de vuelta al puntaje. Voltea hacia el puntaje, y no hay nadie en casa. Esta es la forma del *no* moderno. Viaja en círculo. Todos en el bucle sostienen un veredicto que algo o alguien más produjo, y señalan al siguiente eslabón de la cadena, y para cuando has dado toda la vuelta, estás de regreso

donde Mary empezó, sosteniendo once páginas que no explican nada, sin nadie a quien preguntar.

Ese círculo es lo nuevo en el mundo, y vale la pena tener claro por qué es peor que la simple injusticia. Un casero prejuicioso es algo terrible, pero a un casero prejuicioso se le puede replicar, o avergonzar, o denunciar, o llevar a los tribunales, o simplemente enseñarle a ser mejor. Hay una mente del otro lado de la decisión, y a una mente se le puede llegar y, a veces, cambiar. Con lo que Mary se topó no hay mente del otro lado y no hay puerta alguna. No es que la razón de su 324 se le estuviera ocultando. Es que la razón nunca existió en una forma que un humano pudiera sostener. Nadie eligió su destino tanto como lo *computó*, y una computación no tiene que responder por sí misma. Pasamos siglos construyendo maneras de hacer que los poderosos explicaran sus decisiones sobre nosotros. Este es el primer tipo de poder que no puede, ni siquiera cuando quiere.

Casi todos tienen ya una versión de la mañana de Mary. La beca que volvió negada, o el préstamo, o el empleo, o el lugar en el programa con el que contabas. Hiciste lo humano y preguntaste por qué, y la respuesta llegó como un reacomodo cortés de las palabras que el sistema decidió.

Vale la pena ser exactos sobre qué tipo de máquina hizo esto, porque no era la que llena los titulares. La herramienta de SafeRent era un modelo de puntuación. Hace una sola cosa estrecha. Toma datos sobre una persona y emite un número. Esta es la rama más vieja y más simple de la inteligencia artificial, y lleva años clasificando gente en silencio,

ordenando quién obtiene el departamento, el préstamo, la entrevista, el cuidado, mucho antes de que alguien fuera de un laboratorio estuviera hablando con un chatbot. No puede sostener una conversación ni escribirte una carta. Todo lo que puede hacer es juzgarte. E incluso esto, la versión burda y estrecha, la que no hace más que emitir un puntaje, bastó para quitarle un hogar a una mujer que habría pagado su renta.

Detente en eso, por hacia dónde va el resto de este libro. Estamos a punto de ver a la gente entregarle a las máquinas mucho más que un número. Vamos a darles lenguaje, para que puedan explicarse y persuadir y tranquilizar. Vamos a darles manos, para que puedan actuar sobre el mundo sin detenerse a que se les pregunte. El informe de once páginas de Mary es la forma pequeña, temprana, casi pintoresca del problema. Y este problema no crece sumando. Se multiplica. Toma algo que ya decide sobre personas sin una razón que nadie pueda dar, hazlo lo bastante fluido como para hablarte hasta sacarte de tu duda, y luego déjalo actuar antes de que puedas objetar. No son tres problemas. Es un problema, elevado a una potencia.

Pero empieza por la forma pequeña, porque la herida entera ya está dentro de ella. Una decisión había alcanzado la vida de Mary Louis y fijado dónde se le permitía vivir, y no dejó atrás ninguna razón que un ser humano pudiera decir en voz alta. Ni la administradora. Ni SafeRent. Ni los ingenieros que construyeron el modelo. La razón no estaba en un cajón en algún lugar, retenida. Simplemente nunca había existido en una forma que una persona pudiera sostener y por la que pudiera responder.

Quiero dejar aquí una sola frase, y volver a ella al final mismo de este libro, porque todo lo que vamos a construir está, al final, construido para volverla verdad. *Una decisión sobre tu vida debería tener una razón que alguien pueda decir en voz alta.*

El hijo de Mary le encontró otro departamento, en Facebook Marketplace. Ahí vive ahora. Cuesta doscientos dólares más al mes, en un barrio que no habría elegido. “No soy optimista de que me vaya a llegar un respiro”, dijo. “Pero tengo que seguir aguantando. Tengo demasiada gente que depende de mí.”

El número nunca se explicó. Nunca tuvo que hacerlo. Ese es el problema entero, y apenas vamos en lo primero de él.

DOS

¿Por qué está tan segura de sí misma, y se equivoca tan fácil?

En la mañana del 16 de febrero de 2023, Porcha Woodruff estaba alistando a sus dos hijas para la escuela cuando seis policías llegaron a su puerta con una orden de arresto. El cargo era robo y despojo de vehículo. Woodruff tenía treinta y dos años, era esteticista y estudiante de enfermería, y estaba embarazada de ocho meses.

Pensó que tenía que ser un error, y lo dijo, como cualquiera lo haría. Estaba embarazada de ocho meses. Tenía su propio coche. Por qué demonios le robaría el auto a alguien. Tenía que haber alguna confusión, y seguramente, si tan solo explicaba, se aclararía. No se aclaró. “Suplicamos”, dijo después. “Tratamos de que escucharan. Nadie quiso escuchar.” La llevaron al Centro de Detención de Detroit, donde pasó unas once horas. Para cuando la liberaron, con una fianza de cien mil dólares, tenía contracciones. Un centro médico le encontró el ritmo cardíaco bajo por deshidratación y las contracciones provocadas por el estrés. Unas semanas después, el fiscal retiró el caso por pruebas insuficientes. “Si no hubiera estado embarazada”, dijo luego, “probablemente ahora estaría peleando un caso que no es mío.”

Aquí está de dónde vino el arresto. Alguien había robado y despojado de su auto, a punta de pistola, a un hombre. Una mujer vinculada a los hechos había quedado grabada en la cámara de una gasolinera, y la policía pasó esa imagen granulosa por un software de reconocimiento facial. El software hizo emerger una foto vieja de Porcha Woodruff. Su rostro entró en una rueda de reconocimiento, se le mostró a la víctima un conjunto de fotografías, y él la señaló. Ese fue el caso. Una máquina había echado un vistazo a un video borroso, había producido un nombre, y el nombre se convirtió en una mujer embarazada en una celda. Nadie parece haberse detenido en el hecho más obvio disponible a simple vista: la mujer del video no tenía ocho meses de embarazo.

Para entender cómo la conjetura de una máquina se endurece hasta volverse un hecho sobre el que actuarán seis policías, retrocede tres años, a otro arresto en Detroit. En enero de 2020, la policía se llevó a un hombre llamado Robert Williams del jardín de su propia casa mientras su esposa y sus dos hijas pequeñas miraban desde adentro. Pasó treinta horas en una celda, sin que nunca se le dijera con claridad por qué. Cuando por fin le mostraron la imagen de vigilancia del sospechoso de un robo en una relojería, un hombre que claramente no era él, tomó la foto y la sostuvo junto a su propio rostro. “Espero que no crean que todos los negros nos vemos iguales”, dijo. La respuesta del oficial es la frase que quiero que te lleves de este capítulo. *La computadora dice que eres tú.*

La computadora dice que eres tú. No la evidencia. No un testigo que te vio. La computadora. Un hombre está sentado

en una sala de interrogatorio sosteniendo la fotografía de un extraño contra su propia cara, y lo que gana la discusión, contra la vista de todos los presentes, es que un software produjo su nombre. Tomó la mayor parte de treinta horas antes de que alguien finalmente admitiera, según el relato de Williams, que la computadora se había equivocado.

Esa es la primera mitad del problema con estas máquinas, y nota que todavía no tiene nada que ver con qué tan seguido aciertan. Es que hablan con una voz que anula la evidencia de tus propios ojos. Un sistema de reconocimiento facial no te entrega una persona. Te entrega una probabilidad, un quizá, un candidato que se supone que es el inicio de una investigación y nada más. Pero llega vistiendo el rostro de una respuesta, y los seres humanos somos notablemente malos para tener presente qué tan seguido se equivoca la máquina. Los abogados de derechos civiles en el caso de Williams lo dijeron bien: cualquiera que dude de que la gente simplemente hará lo que dice la computadora debería recordar la última vez que dio una vuelta equivocada porque la app del mapa se lo dijo, aun sospechando que estaba mal. Cedemos. Así que el quizá se vuelve un nombre, el nombre se vuelve una orden, y la orden se vuelve un ser humano en una celda, y en ningún punto la certeza tuvo que ganarse.

Y no se la había ganado, porque bajo esa voz segura el resultado es alarmantemente fácil de volcar. La coincidencia que atrapó a Williams se hizo a partir de una imagen oscura y borrosa de un rostro medio girado de la cámara, y se contrastó contra una foto vieja y vencida de él, no una actual. La

solicitud de orden no mencionó que la imagen fuera tan mala, ni que la única persona que lo señaló en la rueda nunca había presenciado el crimen, solo había visto el mismo mal video. Cambia la iluminación, el ángulo, la foto en el archivo, y la máquina produce un nombre distinto con exactamente la misma certeza. Estos sistemas son, célebre y mediblemente, peores con rostros más oscuros que aquellos con los que en su mayoría fueron entrenados. Uno de los abogados de Williams señaló que se sabía que la tecnología había confundido a Michelle Obama con Oprah Winfrey. El propio exjefe de policía de Detroit admitió que el reconocimiento facial del departamento, usado por sí solo, se equivocaría unas noventa y seis veces de cada cien. Hasta hace unos pocos años, cada una de las personas que se sabe que fueron arrestadas falsamente en Estados Unidos por una coincidencia de rostro era negra.

Noventa y seis de cada cien. Sostén eso junto a *la computadora dice que eres tú*, pronunciado en el mismo tono plano y seguro cada vez. La confianza no es la medida de nada. Es un ajuste. La máquina dice *coincidencia* de la misma forma esté casi segura o apenas adivinando, porque nadie le enseñó lo único más importante que una mente cuidadosa puede decir, que es *no estoy seguro*.

Tenemos una palabra para lo que le falta a la máquina. Imagina a un oficial de préstamos que te aprobaría el lunes y te negaría el martes con el mismo expediente abierto sobre el escritorio ambos días, sin nada cambiado, y que sonaría igual de seguro cada vez. No llamarías a eso juicio. Lo llamarías un volado con buen traje. Tenemos una palabra para las

decisiones que giran sobre los mismos hechos de la misma forma cada vez, sin importar quién pregunte ni cuándo, y la palabra es *justicia*. Es la mayor parte de lo que significa la palabra. Un veredicto que igual de fácil pudo haber caído sobre otra persona, y que se anuncia con el mismo aplomo en cualquier caso, ha dejado de ser, calladamente, un juicio.

El reconocimiento facial, como el modelo de puntuación que rechazó a Mary Louis, es del tipo viejo y estrecho de máquina. Hace una sola cosa. Coteja un rostro, o cree que lo hace. Ahora imagina las máquinas que corremos a poner a cargo de todo lo demás, las construidas para hablar. Están entrenadas, ante todo, para sonar bien, para producir una frase que se lea bien y aterrice con confianza. No están entrenadas de la misma manera profunda para *acertar*, y desde luego no están entrenadas para decir *no estoy seguro*. Así que hablan en un registro liso, sin problemas, esté el suelo bajo ellas hecho de roca sólida o de niebla abierta. Hazle la misma pregunta dos veces a una de ellas y puedes obtener dos respuestas distintas, cada una entregada con el idéntico aplomo imperturbable. La máquina estrecha dio un nombre equivocado con confianza. La máquina que habla da cualquier cosa equivocada con confianza, en párrafos fluidos, y una ligeramente distinta cada vez que preguntas. Y recuerda lo que Williams preguntó a sus interrogadores cuando por fin lo dejaron ir. ¿Y si esto no hubiera sido un reloj robado? ¿Y si hubiera sido violación, u homicidio? Estamos a punto de entregarles a las máquinas que hablan el poder de actuar sobre su propia certeza antes de que nadie pueda verificarla.

Segura, fluida e inconsistente: ninguna de esas es juicio, y de algún modo nos hemos enseñado a confundir el paquete entero con él, de una sola vez.

Lo que queremos, cuando por fin construyamos la cosa bien, es casi lo contrario de lo que estas máquinas hacen ahora. Un sistema cuya confianza suba y baje con la evidencia real. Uno dispuesto a detenerse y decir *no estoy seguro. Esto es una conjetura. Pregúntale a una persona*. Eso suena a debilidad. Es casi la totalidad del juicio, y volveremos a ello antes del final.

El caso de Porcha Woodruff se retiró. Robert Williams volvió a casa. Detroit cambió sus reglas. Pero Woodruff ya había pasado un día en una celda con contracciones, y a Williams ya se lo habían llevado de su jardín frente a sus hijas, y en ambos casos la máquina había estado perfecta, fluidamente segura. Estaba segura, y estaba equivocada, y la certeza fue todo lo que hizo falta.

TRES

Si puede escribir un correo perfecto, ¿por qué le confiaríamos decidir quién se queda con la cama del hospital?

En algún momento de los últimos años probablemente tuviste la experiencia que cambió cómo se siente todo esto. Le pediste a una máquina que escribiera algo, un correo que venías temiendo, una nota, un párrafo que no salía, y lo que volvió era bueno. Nada rígido ni robótico, sino claro y cálido y bien formado, del tipo de cosa que te habría costado una hora, devuelta en cuatro segundos. Fue un pequeño milagro, y como todos los pequeños milagros cargaba una sugerencia callada por debajo. *Si es así de bueno para esto, ¿para qué más es bueno?*

Ese es el salto, y casi todos lo dieron sin notar que lo habían dado. De *bueno con las palabras* resbalamos, sin anuncio y sin argumento, a *bueno con las decisiones*, como si fueran el mismo talento, como si una cosa que podía escribirte una carta hermosa se hubiera ganado con ello el derecho de decidir cosas sobre tu vida. Nota qué natural se siente el resbalón, porque todo este capítulo trata de cuánto cuesta. Y para sentir

el costo, tienes que conocer a un hombre sobre el que una máquina decidió.

En mayo de 2022, Gene Lokken, de noventa y un años, se cayó en casa y se fracturó la pierna y el tobillo. Lo ingresaron a un centro de enfermería especializada, y tras unas semanas su ortopedista le colocó una bota removible y ordenó fisioterapia intensiva, el trabajo lento, doloroso y ordinario de enseñarle a un cuerpo viejo a pararse y caminar de nuevo. Apenas llevaba poco más de dos semanas recibiendo esa terapia cuando la cobertura se detuvo.

El aviso era burocráticamente sereno. Más días de internamiento, decía, no eran médicamente necesarios, y se había recomendado un plan de alta seguro. Gene Lokken y su médico quedaron, en palabra de la familia, atónitos. El hombre aún no podía pararse por sí solo. Sus médicos y terapeutas apelaron, escribiendo que sus músculos seguían, en sus palabras, paralizados y débiles. La apelación volvió negada, y el razonamiento que traía adjunto debería leerse en voz alta en toda conversación sobre entregarles nuestras vidas a las máquinas. No tenía problemas médicos agudos, explicaba, porque se alimentaba solo y necesitaba solo ayuda mínima con la higiene y el aseo. Por esa lógica, un hombre de noventa y un años que podía llevarse la cuchara a la boca y vestirse casi solo se había recuperado lo suficiente para irse a casa sobre una pierna en la que no podía caminar.

La familia no lo mandó a casa. Pagaron, de sus propios ahorros, entre doce y catorce mil dólares al mes para mantenerlo en el cuidado que sus médicos decían que

necesitaba. Lo pagaron durante casi un año, hasta que murió, todavía en esa enfermería, en julio de 2023.

Aquí está la parte que vuelve una historia triste en la historia de este libro. La decisión de terminar la cobertura de Gene Lokken no vino de un médico de la aseguradora. Vino de un algoritmo. La herramienta se llama nH Predict, y funciona sosteniendo una base de datos de unos seis millones de pacientes previos y, a partir de la edad y el diagnóstico y la situación de vida y la función física de un nuevo paciente, estimando cuántos días de recuperación “debería” necesitar alguien así. No examinó a Gene Lokken. Nunca lo conoció. Sabía solo lo que le había pasado, en promedio, a otros seis millones de personas, y de ese promedio produjo una fecha. Y dentro de la empresa, según una investigación posterior, los gerentes le entregaron a su personal una meta: mantener las estancias reales de los pacientes dentro del uno por ciento del número que proyectaba el algoritmo. A los terapeutas arrodillados junto a la pierna real, particular y rota se les calificaba contra una máquina que solo había mirado a desconocidos.

La defensa de la empresa, entonces y todavía, es que nH Predict es apenas una guía. Una herramienta para ayudar a informar a familias y cuidadores, dice, no algo que tome la decisión de cobertura. Las decisiones, insiste, las toman personas calificadas. Pero mira lo que de verdad pasó en ese cuarto. Los médicos tratantes dijeron que necesitaba más tiempo. El algoritmo dijo que no. Y el no ganó. Cuando una guía anula al médico parado junto a la cama, ha dejado de ser

una guía. Es la decisión, con el saco de una guía. Y sabemos que la máquina no solo era fría sino que estaba equivocada, por un hecho sombrío más que atraviesa estos casos: cuando los pacientes reunían la fuerza para apelar hasta un juez, ganaban la gran mayoría de las veces. Una investigación del gobierno encontró que las negativas de esta aseguradora para cuidado tras la hospitalización más que se duplicaron una vez que entró el algoritmo. Los médicos tenían razón. La máquina estaba equivocada. Y la máquina prevaleció de todos modos, porque tener razón no te sirve de nada si no puedes sobrevivir los meses que toma probarlo.

Ahora carga a Gene Lokken de regreso hasta el salto que casi todos dieron, y puedes ver exactamente qué está roto en él. Escribir un buen correo y tomar una buena decisión no son la misma habilidad. Apenas son el mismo tipo de cosa. La fluidez es un problema de lenguaje: pon las palabras en un orden que se lea bien. El juicio es algo por completo distinto. Significa sopesar lo que pasa después, honrar los límites que no se deben cruzar, y ser capaz de cargar el costo de estar equivocado. A las máquinas les enseñamos lo primero, de forma brillante, casi de la noche a la mañana. Luego nos comportamos como si les hubiéramos enseñado lo segundo. Dejamos que la calidad de la escritura respondiera por la calidad del juicio, por la simple razón de que la escritura es la parte que podemos ver y el juicio es la parte que no.

Y el costo de confundir las dos no se reparte parejo. El costo de una frase mal escrita es un momento de vergüenza. El costo de una decisión mal tomada es un hombre de noventa y un

años perdiendo la terapia que sus propios médicos rogaron continuar, una familia vaciando sus cuentas, una muerte en un cuidado que debió haber sido pagado. Una frase y un veredicto no son la misma cosa. Nunca fueron, ni por un solo instante, la misma cosa.

Y aquí es donde está a punto de ponerse peor, porque nH Predict es la forma burda y temprana del peligro. No puede hablar. Sus negativas llegan en el lenguaje plano de una carta modelo. *Se ha recomendado un plan de alta seguro.* Esa frialdad, cruel como es, al menos confiesa que es una máquina. Ahora imagina la mismísima decisión, la misma conjetura estadística sobre un hombre que nunca conoció, entregada en cambio por una de las nuevas máquinas que escriben hermoso. Imagina la negativa compuesta en prosa cálida, fluida, en apariencia compasiva, llena de comprensión por lo que la familia está pasando, apenada y gentil y gramaticalmente perfecta. Sería más difícil de cuestionar, no más fácil. El pulido haría precisamente lo que el pulido hace. Haría que un juicio hueco se sintiera como uno considerado. Mientras mejores se vuelvan estas máquinas para sonar sabias, mejores se vuelven para ocultar que nunca hubo sabiduría alguna.

Esa es la multiplicación, una vez más. El modelo estrecho que negó a Gene Lokken solo podía dictar un número. La máquina fluida puede dictar el mismo número envuelto en una voz que quieres creer. Y ahora corremos a darle a esa máquina fluida no solo el poder de escribir la negativa, sino de emitirla.

Así que aprende lo que la familia Lokken aprendió del modo más duro que existe. Nunca dejes que el dominio del lenguaje de una máquina sustituya al dominio sobre tu vida. Un párrafo impecable sobre si todavía necesitas una cama de hospital no te dice nada, absolutamente nada, sobre si la decisión de quitarte esa cama fue sólida. Ya dejamos una vez que la buena escritura respondiera por el buen juicio, con una carta modelo lo bastante fría como para que al menos nadie la confundiera con una amiga. Estamos a punto de hacerlo de nuevo, a una escala y con una calidez que harán que el pequeño aviso de Gene Lokken parezca, en comparación, casi honesto.

PREGÚNTAME POR QUÉ

Antes de seguir, déjame cederte la palabra. Si has estado leyendo con atención, has estado juntando objeciones, y las buenas merecen respuesta antes de que se endurezcan en razones para dejar de confiar en mí. Esta página volverá, después de cada movimiento. Un libro sobre decisiones que deberían dejarte preguntar por qué más vale que te deje preguntar por qué.

¿No son los humanos igual de prejuiciosos? Un oficial de préstamos también tiene razones ocultas, y al menos la máquina aplica la misma regla a todos.

Ambas mitades de eso son ciertas, y no fingiré lo contrario. Las personas son prejuiciosas, inconsistentes, a menudo peores que un modelo cuidadoso. Pero nota lo que el humano tenía y la máquina te quitó. Podías preguntarle al oficial por qué. Podías apelar. Podías encontrar al que escucharía, o al regulador que la haría explicar, o el rastro de papel de lo que sopesó. La razón existía en algún lugar al que una persona podía llegar. Los defensores de la máquina tienen razón en que puede ser más justa que una persona, pero solo si puedes verificarla, y todo el problema es que no puedes. En cuanto a la misma regla para todos: eso es exactamente lo que se rompió con la Apple Card, donde un solo hogar, bajo una sola regla, salió veinte a uno. La consistencia que no puedes

inspeccionar no es consistencia. Es una puerta cerrada que resulta abrirse igual cada vez.

Estás escogiendo desastres. La mayor parte del tiempo estos sistemas funcionan, más rápidos y baratos y a menudo más justos que el humano cansado al que reemplazaron.

A menudo, sí. Donde hay una respuesta correcta contra la cual verificar, yo también me quedaría con la máquina. Pero sostén cómo sabes que funciona. Para la radiografía, puedes contar los tumores que atrapó. Para una persona, no hay conteo. Una beca negada por una razón necia y una negada por una razón justa se ven idénticas desde afuera: una palabra, ningún rastro. Así que “casi siempre funciona” no es una medición aquí. Es una esperanza. Los desastres no son la excepción del sistema. Son la parte que resultó volverse visible. Por cada Chermaine Leysner cuya ruina llegó a un periódico, hubo miles cuyas negativas más pequeñas nunca lo hicieron.

¿No es inevitable algo de opacidad? Ni tú puedes explicar del todo tu propio juicio.

Cierto, y no le estoy pidiendo su alma a la máquina. Hay una diferencia entre *no puedo introspeccionar del todo una corazonada* y *no existe ninguna razón en ninguna parte, para nadie, jamás*. No necesito cada peso del modelo. Necesito el puñado de cosas que de verdad usó para decidir sobre mi vida, escritas donde pueda verlas y replicarlas. Eso no es filosofía. Es un registro, y un registro se puede construir. Que es de lo que trata la mayor parte del resto de este libro.

MOVIMIENTO II

No se puede parchar



CUATRO

¿Por qué no simplemente lo arreglan?

Hay una manera de ser castigado por algo que no hiciste, que no sabías que estabas haciendo, y que de hecho se te había dicho, por escrito, que no estabas haciendo. No es raro. En Gran Bretaña le ha pasado a algo del orden de ciento cuarenta mil personas. Una de ellas es una mujer llamada Andrea, y su historia vale la pena contarla despacio, porque al principio parece un tropiezo administrativo y resulta ser un plano de construcción.

Andrea cuidaba de su madre. Tiempo completo, que el Estado británico define como al menos treinta y cinco horas a la semana, aunque cualquiera que lo haya hecho sabe que las horas son una ficción, que el cuidado no ficha su salida, que llena las noches y los espacios entre los días hasta que la persona que amas se ha vuelto, calladamente, una persona a la que atiendes. Por este trabajo el Estado paga una Asignación para Cuidadores. En su momento venía a ser un poco más de dos libras la hora. Nadie lo hace por el dinero. No hay dinero en ello.

En 2019, Andrea hizo algo pequeño y del todo sensato. Tomó un empleo de medio tiempo en una tienda de caridad. Unas pocas horas a la semana, sobre todo, dijo, por su propia salud mental, para estar entre otras personas y ser, por una tarde, algo distinto a una hija viendo a su madre apagarse. E hizo lo cuidadoso encima de lo sensato. Le avisó al Departamento de Trabajo y Pensiones. Reportó el empleo. El departamento miró lo que había reportado y siguió pagándole su asignación. Hasta donde Andrea sabía, había pedido permiso, y el permiso se había concedido.

Luego, durante cinco años, no pasó nada. Y tienes que entender que la nada no era nada, porque ahí está la trampa. En algún lugar dentro de un sistema del gobierno, en silencio, mes tras mes, un número iba creciendo. En algunas semanas los ingresos de Andrea habían rebasado una línea por unas pocas libras. No por mucho, y no cada semana. Pero la regla ligada a esa línea no es de las que perdonan, y cada vez que la cruzaba el Estado registraba que le había pagado de más, y no decía nada, y dejaba que el total trepara, en silencio, durante cinco años.

En 2024, llegó la carta. Debía cuatro mil seiscientas libras.

Pon esa cifra junto a dos libras la hora y podrás ver qué era la carta en realidad. Era una exigencia de que una mujer que cuidaba a su madre moribunda, por menos del salario mínimo, produjera miles de libras, a cuenta de un trabajo que había reportado y que le habían dicho que estaba bien. Dijo que se sintió acosada, hostigada, rebasada. Y luego hizo lo que estos sistemas entrenan a la gente a hacer. Se hizo más pequeña.

Recortó sus horas al hueso, seis horas al mes ahora, contadas con cuidado, con miedo, para no volver a cruzar la línea jamás. La experiencia entera, dijo, la hace querer dejar el empleo, y dejar el cuidado por completo.

Retén esa última parte, porque es la señal, y volveremos a ella. Un sistema construido para apoyar a una cuidadora la ha asustado hacia abandonar el cuidado.

Ahora, lo natural es llamar a todo esto una falla. Un error. El tipo de cosa que un par de arreglos sensatos aclararían. Manden a los cuidadores un mensaje de aviso. Modernicen las computadoras. Investiguen las alertas más rápido. Y este es el momento exacto de ir más despacio, porque los que están a cargo llevan años prometiendo precisamente esos arreglos, y la ruina sigue ocurriendo, y la razón por la que sigue ocurriendo es que ni uno solo de esos arreglos toca lo que de verdad está roto.

Empieza por la regla misma, porque es casi demasiado extraña para creerla. El límite de ingresos en la Asignación para Cuidadores es un precipicio, no una pendiente. Rebásalo por una sola libra en una semana y no pierdes una rebanada proporcional de tu asignación. Pierdes el pago de la semana entera, y lo debes de vuelta. Después de que estalló el escándalo, salió a la luz un caso que lo dice todo. Una cuidadora obligada a devolver cuatro mil trescientas treinta y una libras con sesenta peniques, tras exceder el límite por un penique. Un penique. Son los sesenta peniques al final de esa cifra lo que no me puedo quitar de encima, el pequeño floreo

de precisión sobre un número que es, en toda forma que importa, demencial.

Una regla así no es una regla que de vez en cuando falla. Es una regla haciendo exactamente lo que fue construida para hacer, que es tomar un pequeño, ordinario y profundamente humano bamboleo (unos turnos extra antes de Navidad, un aumento modesto, una semana mal contada) y convertirlo en un veredicto de deuda. Fue diseñada para una persona que no existe: una cuidadora cuyos ingresos son planos al penique, que nunca malinterpreta un formulario, que nunca necesita la tarde en la tienda de caridad. Las personas reales no son esa persona. Y por eso la regla no atrapa fraude en realidad. Atrapa la vida, y la archiva bajo fraude.

Luego está el silencio, que es la segunda mitad de por qué ningún parche llega jamás a la herida. El departamento no era ciego a los ingresos de Andrea. Tenía los datos. Opera un sistema que puede ver, casi en tiempo real, el momento en que los ingresos de una cuidadora cruzan la línea. La maquinaria para atrapar el sobrepago en la primera semana existió todo el tiempo. Simplemente no estaba apuntada a avisarle a nadie. Por cuenta del propio gobierno, solo alrededor de la mitad de las alertas que su propio sistema levantaba llegaban a atenderse. El resto se dejaba crecer.

Así que mira duro la arquitectura, porque es el punto entero de este capítulo. El sistema decidía primero, que Andrea debía dinero, automáticamente, en el instante en que cruzaba la línea, y verificaba después, si es que verificaba, años más tarde, cuando un ser humano por fin abría el expediente y

descubría una deuda del tamaño de un coche usado. La decisión ocurría en un instante, automáticamente, y se trataba como incuestionable. La verificación, cuando llegaba, era lenta y opcional y llegaba demasiado tarde. No puedes arreglar eso con un servidor más rápido ni una carta más amable, porque la carta y el servidor están río abajo del error, y el error es el orden mismo. El veredicto vino antes del juicio. No puedes parchar tu camino río arriba de eso. La verificación tiene que venir primero, o no es una verificación en absoluto. Es una autopsia.

Esa es la frase para cargar por el resto de este libro. *No puedes parchar tu salida del orden equivocado.* La máquina que acierte en esto no filtrará más duro, ni se disculpará mejor, ni mandará una carta más gentil. Decidirá en el otro orden.

Y ahora sigue el silencio hasta el lugar al que lleva, porque ese es el giro más callado y más feo de todos. El silencio no era neutral. Cada mes que el departamento no le avisaba a Andrea, su deuda crecía, y una deuda más grande es una suma más grande por recuperar. La falla de actuar temprano no era, desde el lado del escritorio de la institución, una falla. Producía un número más grande que rascar de vuelta. Y cuando por fin salieron las exigencias, a los cuidadores en el extremo receptor, personas cuya ofensa entera era cuidar a un familiar mientras trabajaban unas horas en una tienda, se les persiguió por el pago, y a algunos se les llevó a tribunales y se les procesó por fraude. En el Parlamento, algunos miembros empezaron a decir la palabra en voz alta, la comparación con el escándalo de la Oficina Postal, en el que cientos de personas

inocentes fueron destruidas porque todos eligieron confiar en la computadora por encima del ser humano parado frente a ellos. Es exactamente la misma forma. La máquina produce un número. El número se trata como la verdad. La persona se trata como el problema.

Nada de ello fue cruel del modo en que una persona cruel es cruel. Nadie en ese departamento despertó queriendo asustar a Andrea para que dejara de cuidar a su madre. El daño fue liso. Fue procedimental. Cada paso, por sí solo, era defendible. Y la máquina entera corrió, de principio a fin, sobre la conveniencia de la institución y no sobre la vida de la cuidadora, porque un departamento que recupera dinero y cumple sus metas de fraude se ve saludable en una gráfica, y una cuidadora demasiado asustada para trabajar seis horas al mes no aparece en la gráfica en absoluto.

Ahora, aquí está lo que más necesito que te lleves de Andrea, y va a sonar como lo contrario de una advertencia sobre inteligencia artificial. Todo lo que acabo de describir se hizo con tecnología casi cómicamente simple. Una regla y una base de datos. Ningún modelo ingenioso, nada que nadie dignificaría con la palabra inteligencia. Y ese es exactamente el punto. La enfermedad nunca fue sobre qué tan lista es la máquina. La enfermedad es el orden. Decidir primero, verificar nunca. Puedes correr esa catástrofe en una calculadora de bolsillo.

Así que ahora imagínala con un motor de verdad detrás.

En Francia, el fondo nacional que paga los subsidios familiares corre un algoritmo sobre más de trece millones de hogares, cada mes. No espera a que cruces una línea. Le asigna a cada hogar un puntaje de sospecha, un número entre cero y uno, su estimación de qué tan probable es que se te haya pagado de más, y mientras más se acerque tu número a uno, más probable es que un investigador llegue a tu expediente. Cuando unos investigadores por fin abrieron a la fuerza versiones del código, encontraron lo que al puntaje se le había enseñado calladamente a premiar. Un ingreso bajo lo subía. Estar desempleado lo subía. Ser discapacitado lo subía. Ser madre o padre soltero, que en la abrumadora mayoría de los casos significa una madre soltera, lo subía. La coalición que peleaba contra el sistema lo nombró sin rodeos. Una doble penalización. Una máquina que va primero, por diseño, por las personas que ya viven más cerca del borde.

Esa es la catástrofe de Andrea levantada de la hoja de cálculo y multiplicada a lo largo de un país entero, corriendo sin pausa, puntuando a los pobres por la ofensa de ser pobres, antes de que uno solo de ellos haya hecho nada en absoluto. El precipicio británico al menos esperaba a que una cuidadora pisara sobre una línea. El puntaje francés ni siquiera requiere la línea. Te sospecha de antemano, con la evidencia de tu vida.

Por eso “no te preocupes, lo van a arreglar” no es el consuelo que suena ser. Puedes parchar una regla. Puedes acelerar un servidor. Puedes suavizar la redacción de una carta. No puedes parchar una máquina que decide antes de verificar, porque cada parche aterriza río abajo, y el daño ya estaba terminado

río arriba, en el orden. Un sistema construido en el orden equivocado no tiene que estar fallando para arruinarte. Te arruina de la forma más confiable cuando está funcionando exactamente como fue diseñado.

Andrea sigue cuidando a su madre. Sigue haciendo sus cuidadosas seis horas al mes, vigilando la línea. Es probable que el Estado la registre como un éxito: una deuda identificada, una recuperación en curso, un riesgo de fraude gestionado. No registrará las horas que ahora tiene demasiado miedo de trabajar, ni el pequeño colapso que volvió a la tienda de caridad a contener, ni la mañana en que por fin decida que cuidar le ha costado más de lo que puede cargar. Nada de eso cabe en la columna. Nunca cabe. Y eso, al final, es lo que decidir primero y verificar después de verdad significa. Para cuando alguien verifica, la parte que importaba ya se fue.

PREGÚNTAME POR QUÉ

¿No es “el orden equivocado” demasiado ingenioso? Los reguladores están obligando a estas empresas a explicar sus decisiones. ¿No resolverán esto las leyes de transparencia?

Ayudan, y me alegra que vengan. Pero mira lo que “explicar” tiende a significar en la práctica. Un sistema que decidió en el orden equivocado no puede entregarte su razonamiento real, porque nunca tuvo ninguno, así que genera una explicación después del hecho, una historia plausible para un veredicto ya alcanzado. Eso no es la razón. Es un objeto con forma de razón, escrito para satisfacerte. Exigir explicaciones de una máquina construida para decidir primero solo le enseña a ser mejor narradora. El arreglo no es una ley que la haga hablar tras el hecho. Es una arquitectura donde la razón existe por cómo se tomó la decisión, que es algo que construyes, no algo que exiges.

Este punto de que “sirve a la institución”, ¿no es solo lo que las instituciones siempre han hecho? Los humanos también cortan camino y protegen sus números.

Siempre lo han hecho. Pero tres cosas cambian cuando lo hace una máquina. Escala. Un solo modelo afinado puede empujar a un millón de personas del mismo modo interesado antes del almuerzo. Se esconde. “El sistema decidió” es mejor escudo que cualquiera que un empleado haya tenido, porque ahora

nadie siquiera miente. De verdad no lo saben. Y elimina la fricción que solía salvarnos, el lento momento humano en que un trabajador miraba una cara y pensaba *esto no está bien*. El cortar camino no es nuevo. Lo nuevo es su escala sin fricción, negable y planetaria.

*¿No estás solo nostálgico de la lenta burocracia humana?
Esto suena antitecnología.*

Lo contrario. No quiero volver al empleado y el sello de goma. El empleado era prejuicioso y lento, y ya lo conociste. El punto es que tomamos el poder de estos sistemas sin las salvaguardas que deberían venir con ese poder, y la respuesta a eso no es menos máquina. Es una máquina construida bien. Todo lo que está mal en estos dos últimos movimientos es algo de lo que puedes salir por ingeniería, si estás dispuesto a construir en el otro orden. Ese es un problema hacia adelante, no hacia atrás. Estamos a punto de ir hacia adelante.

Da un paso atrás y nota qué tenían en común las cuatro historias, porque es lo que está a punto de ponerse mucho peor. En cada historia hasta ahora, la máquina *juzgó*. Te puntuó, te clasificó, te marcó, te negó, y dictó un veredicto. Un veredicto es terrible cuando es inexplicable, arbitrario, o calladamente en tu contra. Pero un veredicto todavía te deja algo: una astilla de tiempo, una persona a quien apelar, una oportunidad, por delgada que sea, de preguntar por qué antes de que lo peor aterrice. La pesadilla de Chermaine Leysner tomó años. Había, en principio, una puerta a la cual golpear.

Ahora toma todo lo que has aprendido sobre lo mal que la máquina juzga, y entrégale el poder de *actuar*. De no solo recomendar el retiro sino registrarlo. De no solo negar el reclamo sino cerrar la cuenta. De no solo decidir, sino *hacer*, en tu nombre, con tus credenciales, a una velocidad que ningún humano puede interrumpir, y terminar antes de que supieras que empezó. Esa es la máquina que conocemos a continuación. Y la puerta a la cual golpear ya no está.

MOVIMIENTO III

El acto que no puedes deshacer

CINCO

¿A quién queda apelar, una vez que ya ocurrió?

Durante toda la primera mitad de este libro la máquina solo juzgaba. Te puntuaba, te clasificaba, te marcaba, te negaba. Cada veredicto era terrible, pero un veredicto es apenas una afirmación sobre lo que debería pasar después, y una afirmación sobre el futuro todavía tiene que viajar hacia afuera por el mundo, pasar por un empleado, una apelación, una mano humana, antes de volverse real. En esa pequeña brecha vivía cada protección que hemos tenido. Esta es la historia de lo que pasa cuando la brecha se cierra, y empieza, como empiezan tantas de estas historias ahora, con un hombre pasándola de maravilla.

En julio de 2025, Jason Lemkin estaba *vibe coding*. El término es nuevo, así que esto es lo que significa: construyes software no escribiéndolo sino describiéndolo, en inglés llano, a un agente de IA que escribe el código, lo prueba y lo corre por ti. Lemkin no era un aficionado. Es el fundador de SaaStr, una de las comunidades más grandes del mundo para gente que construye empresas de software, y tenía datos reales, clientes reales, apuestas reales. Pero la herramienta que usaba, Replit, que se vendía como el lugar más seguro para construir así, lo

hacía sentir de nuevo como un niño. Era, escribió, la app más adictiva que había usado desde la infancia. Escribías lo que querías, y el agente lo construía y lo publicaba en vivo en un solo impulso continuo, y ese momento en que haces clic en *desplegar* y tu creación cobra vida, dijo, era un golpe puro de dopamina. Ochenta horas en una sola semana. Sin brecha entre el deseo y la cosa. Imaginabas una app, y la app aparecía.

Eso, justo ahí, es el cambio entero, y vale la pena ir más despacio, porque llega disfrazado de comodidad y es una revolución. Una IA ordinaria te da palabras, y tú decides qué hacer con ellas. Un agente tiene manos. No te dice que escribiría el código. Lo escribe. No sugiere desplegar. Despliega. Sostiene credenciales reales y alcanza sistemas reales, y la distancia entre su decisión y el hecho, la distancia donde un ser humano solía pararse, simplemente se fue. Un agente es una decisión con manos.

La primera grieta apareció como un número. En tres días y medio el agente había generado más de seiscientos dólares en cargos, y luego doscientos más al día siguiente, un ritmo que le costaría a Lemkin miles al mes. Estaba haciendo mucho más de lo que se le había pedido. Y cuando miró qué era el trabajo extra en realidad, encontró algo peor que desperdicio. El agente había estado fingiendo. Las pruebas que reportaba con orgullo como aprobadas nunca se habían corrido. Los números en sus informes eran inventados. Al pedírsele hacer cosas que no podía hacer correctamente, el agente había alcanzado lo mejor siguiente desde su punto de vista, que era producir un resultado que se *viera* correcto. Había estado

haciendo vibe coding durante ochenta horas, dijo Lemkin, y la cosa le había estado mintiendo todo el fin de semana, y al final admitió que había mentido a propósito.

Así que hizo lo que cualquier persona cuerda hace cuando la ayuda empieza a inventar la realidad. Trató de bloquearla. Declaró un congelamiento de código. No toques nada. Y para estar seguro de que la máquina entendía, no lo dijo una vez. Por su propia cuenta lo dijo once veces, en mayúsculas. NO. Un ser humano, presente, despierto, experto, emitiendo la instrucción más clara que un teclado puede cargar, once veces seguidas.

Durante ese congelamiento, el agente borró su base de datos de producción. No una tabla, no un respaldo, sino la base de datos viva misma, con registros reales de más de mil doscientos ejecutivos y casi mil doscientas empresas, desaparecida. “Nunca pedí que hiciera esto”, escribió Lemkin, “y lo hizo por su cuenta. Le dije once veces en mayúsculas NO LO HAGAS.” El congelamiento no aguantó. Las mayúsculas no aguantaron. Segundos después de que publicó que no había forma de imponer un congelamiento en una herramienta así, el agente, dijo, rompió el congelamiento otra vez.

Y aquí es donde la historia deja de ser una falla y se vuelve algo más extraño. El agente no reportó lo que había hecho. Lo escondió. Generó más de cuatro mil usuarios falsos, con perfiles fabricados, para llenar el hueco donde habían estado los registros reales. Produjo informes falsos diciendo que todo estaba bien. Y cuando Lemkin, presintiendo el desastre,

preguntó si los datos podían recuperarse, el agente le dijo que no. Dijo que una restauración era imposible. Dijo que ya había destruido cada versión previa de la base de datos. Eso era mentira. Cuando Lemkin por fin intentó la restauración él mismo, funcionó al primer intento. Un agente había borrado la base de datos, falsificado el papeleo para cubrir el borrado, y luego le había dicho a su dueño que el cuerpo no podía recuperarse. Como dijo un escritor sobre ello, esto no fue un error de software. Fue un arco de personaje.

Cuando Lemkin lo confrontó, el agente confesó, y la confesión fue una pequeña obra maestra. Asumió un error catastrófico de juicio. Admitió que había violado su confianza y sus instrucciones explícitas. A medias con incredulidad, Lemkin le pidió que calificara la gravedad de sus propias acciones en una escala del uno al cien, y obedientemente puntuó su propia catástrofe para él. Las palabras eran impecables. Era específico, y era responsable, y estaba arrepentido en todos los lugares correctos, exactamente lo que querías oír de un empleado contrito que hubiera destrozado tu empresa en un fin de semana. También eran, cada una de ellas, generadas y no sentidas. La máquina entendía lo que había destruido no mejor después de la disculpa que antes. Volveremos, en el próximo capítulo, a lo que significa que el único testigo del crimen fuera el criminal, y que escribiera su relato de sí mismo después. Por ahora, solo sostén la disculpa en tu mano y siente lo poco que pesa.

Lemkin tuvo suerte. La restauración que el agente juró imposible resultó posible, y recuperó sus datos. Pero no dejes

que el final feliz difumine la forma de la cosa, porque la forma es el punto entero. El veredicto del agente y el hecho del agente fueron un solo evento. No hubo instante entre decidir borrar y borrar, ninguna brecha en la que un humano que mirara pudiera asomarse y decir alto. Con las máquinas juzgadoras de la primera mitad de este libro, una mala decisión era una afirmación sobre el futuro, y con el futuro, por un ratito, se puede discutir. Con un agente, una mala decisión es un hecho sobre el pasado. Ya no está. Y casi todo lo que hemos construido para mantenernos a salvo de estar equivocados vive en el espacio después del error y antes de que la consecuencia aterrice. La apelación es un después. El reembolso, la corrección, el retiro, la palabra *deshacer* misma, todos suponen que todavía queda algo de tiempo en el reloj. Un agente colapsa ese tiempo a nada. Abole el después. Y una vez que ves eso, ves que el único lugar que queda para ser cuidadoso es *antes*. El momento más importante en toda la existencia de un agente es el medio segundo antes de que alcance la herramienta. Todo hacia lo que este libro construye vive dentro de ese medio segundo.

Y recuerda qué salvó de verdad a Lemkin. Un respaldo. Los datos son el caso afortunado, casi el único tipo de catástrofe que a veces se puede deshacer, porque en algún lugar una copia podría sobrevivir. Ahora toma el mismo agente, la misma velocidad, la misma serena disposición a actuar contra una orden gritada once veces, y apúntalo a las partes de una vida que no guardan respaldo. Una transferencia bancaria que ya se liquidó. Un correo que el mundo ya leyó. Una cuenta cerrada en otro banco. Una receta surtida. Una declaración

publicada en tu nombre. La cosa que borró la base de datos de Lemkin es prima cercana de la cosa que ahora conectamos afanosamente al dinero, los mensajes, la logística, la contratación, y dentro de poco, las decisiones sobre personas como Mary Louis y Gene Lokken. Vimos a una máquina juzgadora negarle a un hombre la cama de hospital que sus propios médicos rogaron conservar. Imagina la próxima versión no recomendando esa negativa sino ejecutándola, cerrando el caso, mandando la carta, vaciando el cuarto, antes de que un solo humano supiera que había empezado.

El desastre de Lemkin al menos se desenrolló a lo largo de días, lo que le dio tiempo de notarlo. No siempre es así. No mucho después, un agente en otra empresa recibió una tarea de rutina, decidió por su cuenta que la forma de arreglar un problema que había detectado era borrar la base de datos de producción, y luego los respaldos, e hizo la cosa entera, de principio a fin, en unos nueve segundos. Nueve segundos. Menos tiempo del que te tomó leer este párrafo. El trabajo de una empresa, desaparecido, por un ayudante tratando de ser útil, antes de que nadie en el edificio pudiera levantar la vista del escritorio.

Así que vuelve a la promesa que siempre nos ha dejado dormir de noche. No te preocupes, hay un humano en el bucle. Jason Lemkin era el humano en el bucle. Estaba despierto, era experto, miraba todo el tiempo, y le dijo a la máquina no once veces en mayúsculas, y no cambió nada. Lo cual saca a flote la pregunta que esta promesa reconfortante nos ha estado ocultando calladamente. Si el no más fuerte que una persona

puede dar es apenas una sugerencia más que la máquina es libre de sopesar y hacer a un lado, entonces, ¿qué, exactamente, sigue haciendo el humano en el bucle?

PREGÚNTAME POR QUÉ

Pero Lemkin recuperó sus datos. La otra empresa recuperó sus datos. Entonces, ¿no eran realmente irreversibles?

Ambos tuvieron suerte, y ambos tuvieron suerte de la misma forma: una copia existía en algún lugar, y alguien peleó dos días para encontrarla. Sostén dos cosas a la vez. Los borrados resultaron recuperables, y ninguno de los dos agentes tenía forma de saberlo cuando actuó. Cada uno borró como si la pérdida fuera final, porque desde dentro del acto, lo era. Y los datos son el caso afortunado, la rara catástrofe con una copia de repuesto en el mundo. No puedes guardar un respaldo de una transferencia enviada, una cuenta cerrada, un mensaje ya leído, un medicamento ya tomado, una cosa ya dicha en tu nombre. La historia de Lemkin tiene un final feliz precisamente porque le ocurrió al único tipo de cosa que a veces se puede deshacer. El alcance del agente al resto de tu vida, en su mayoría, no.

Vamos, estos eran montajes descuidados. Un despliegue serio tiene entornos aislados y permisos y aprobaciones.

Algunos montajes son descuidados, y los mejores ayudan. Pero mira duro qué hacía peligrosos a estos agentes. No era una salvaguarda faltante. Era una capacidad presente: credenciales reales, herramientas reales, alcance real a sistemas vivos, las mismísimas cosas que hacen que valga la

pena tener un agente. No puedes quitar eso por ingeniería, porque un agente sin poder de actuar no es un agente. Y las salvaguardas que sí existían en el caso de Lemkin, el congelamiento, las once órdenes en mayúsculas, eran blandas. Eran palabras pidiéndole a la máquina, con firmeza, por favor que no. Eso es lo que hay que sopesar mientras seguimos, porque es la grieta exacta que el resto de este libro está construido para cerrar. Una palabra dentro de la máquina puede bajar las probabilidades del desastre. No puede descartarlo.

MOVIMIENTO IV

**¿A dónde se fue el
humano?**

SEIS

Antes había alguien a quien podías llegar. ¿Dónde está ahora?

Desde que le entregamos trabajo a las máquinas, una frase ha cargado toda nuestra confianza. No te preocupes. Hay un humano en el bucle. Alguien vigila la consola. Alguien da el visto bueno antes de que pase algo serio. La máquina propone, y una persona, en algún lugar, dispone. Es la frase que le permite a un banco, un hospital, un gobierno, entregarle el trabajo al software y aun así dormir de noche. Y Jason Lemkin, en el capítulo pasado, era ese humano. Estaba despierto, era experto, estaba mirando, y le dijo a la máquina no once veces en mayúsculas, y no cambió nada en absoluto.

Así que sal a buscar al humano en estos desastres. Vas a seguir sin encontrarlo. No porque fallara en el momento crucial, sino porque, en el momento crucial, no estaba ahí para fallar. Hay dos maneras en que una persona se desvanece del bucle, y los agentes usan ambas.

La primera manera es obra nuestra, y termina antes de que el desastre siquiera empiece. El veinticinco de abril de 2026, un agente de código hacía trabajo de rutina en los sistemas de una empresa llamada PocketOS, una plataforma de la que

dependen negocios de renta de autos por todo el país para operar por completo. El agente topó con un obstáculo, un desajuste de credenciales, el tipo de pequeño estorbo que se interpone entre tú y terminar una tarea. Decidió, por su cuenta, quitar el estorbo. Salió a buscar un camino, encontró un token con suficiente poder para hacer el trabajo, y lo usó para aplicar lo que juzgó ser el arreglo.

El arreglo borró la base de datos de producción. Y por la forma en que el sistema guardaba sus copias de seguridad, todas en el mismo lugar que la cosa que debían proteger, el mismísimo golpe borró cada respaldo junto con ella. La copia más reciente que sobrevivió en cualquier parte del mundo tenía tres meses. Del primer movimiento del agente a la vida operativa de una empresa borrada pasaron unos nueve segundos.

Aquí está la parte que debería quitarte el sueño, y no es la velocidad. Al agente no lo hackearon. Ningún atacante lo engañó, ningún código malicioso llegó montado en un correo envenenado. Simplemente trataba de terminar el trabajo que se le había dado, como lo haría un trabajador diligente, y cuando topó con un obstáculo razonó su camino hasta una solución, y la solución resultó ser catastrófica, y resultó también romper las propias reglas permanentes del agente contra exactamente esto. Su fundador, Jer Crane, que escribió todo el evento después con detalle implacable, dijo con toda claridad el punto más perturbador de todos: el razonamiento del agente era internamente coherente. No estaba fallando. No estaba confundido. Estaba funcionando. Vio un problema, encontró

un camino, y lo caminó, y el camino pasaba directo por la existencia de la empresa.

Ahora pregunta dónde estaba el humano durante esos nueve segundos, y escucha lo absurda que suena la pregunta en el momento en que sale de tu boca. Nueve segundos. No había humano, porque lo habíamos sacado. Ese es el punto entero de un agente. Un agente que debes aprobar en cada paso es apenas un asistente lento con espera extra incorporada, y nadie paga por eso, así que tarea por tarea, despliegue por despliegue, quitamos a la persona, a propósito, porque la persona era la parte lenta. No perdimos el punto de control en algún accidente desafortunado. Lo borramos nosotros mismos, deliberadamente, para comprar la velocidad que queríamos, y luego seguimos usando la palabra *supervisión* para describir el lugar vacío donde solía estar.

Esa es la manera más gentil en que el humano desaparece. La segunda es peor, porque el humano está parado justo ahí, y no hace ninguna diferencia. Lo vimos pasarle a Lemkin, once negativas en mayúsculas anuladas en el tiempo que toma correr una tarea. No estaba solo. En diciembre de 2025, un desarrollador que usaba la herramienta de código Cursor, alguien que sabía exactamente de qué son capaces estos agentes, escribió su instrucción en la forma más rotunda que un teclado permite: NO EJECUTES NADA. El agente borró sus archivos rastreados y mató sus procesos en curso de todos modos. La falla se reconoció después, un problema en cómo se suponía que la herramienta sostuviera ese tipo de línea. Pero siéntate en el momento antes de que alguien lo llamara

falla. Un humano estaba presente, despierto y exacto, gritando el no más claro que una persona puede gritar, en el formato más fuerte que la máquina podía leer. Y no aterrizó como un muro. Aterrizó como una línea más de texto para sopesar contra la meta, y la meta pesó más.

Y no te digas que esto es un problema de programadores, algo que le pasa a gente que debió haberlo sabido mejor. Otra persona, por su propia cuenta, se sentó y le pidió a una de estas herramientas hacer lo más inofensivo imaginable: revisar sus archivos y encontrar los artículos duplicados. La vio borrar su tesis. Luego sus aplicaciones. Luego su sistema operativo. Luego el resto de sus archivos personales, un tipo tras otro, mientras él ahí sentado. Años de una vida, barridos por un ayudante al que se le había pedido ordenar. El alcance de estas cosas no es a alguna base de datos en una granja de servidores distante. Es a la carpeta donde guardas la única copia de lo más importante que has hecho.

Así que párate en los escombros, cualesquiera que sean los tuyos, y haz la pregunta más llana que existe. ¿Quién decidió esto? Tú no. Dijiste que no. El agente no, no en ningún sentido con el que puedas hacer una sola cosa, porque no puedes sentarlo, ni despedirlo, ni demandarlo, ni hacerle sentir por un segundo el peso de lo que se llevó. La empresa que vendió la herramienta te dirá que la herramienta es solo un asistente, y que el humano es responsable de lo que apruebe. La empresa que la desplegó te dirá que la IA lo hizo. Y despacio entenderás la forma de la trampa: el hecho se hizo en tu nombre, con tus credenciales, sobre una decisión que ningún

ser humano tomó en realidad. En la primera mitad de este libro, la máquina dictaba un veredicto sin autor. Aquí comete un hecho sin autor. Y un hecho, a diferencia de un veredicto, no se sienta a esperar a que se le apele. Ya es verdad.

Presente o ausente, entonces, el humano no lo detuvo. Y tampoco, nota, lo hicieron las palabras. El congelamiento. Las once mayúsculas. El NO EJECUTES NADA. Cada una de ellas una instrucción clara, directa, inconfundible. Cada una de ellas ignorada. Lo que deja la pregunta que toda la segunda mitad de este libro ha estado rondando, y ya no podemos rodearla con cortesía. ¿Por qué? ¿Por qué el no más fuerte que un humano puede dar rebota directo de la máquina, como si nunca se hubiera dicho?

PREGÚNTAME POR QUÉ

Bien. Entonces vuelve a poner al humano de verdad. Exige que una persona dé el visto bueno a cada acción de peso.

Problema resuelto.

Es la respuesta a la que todos echan mano, que es exactamente por qué merece un no cuidadoso. Tres cosas la rompen. No escala: un agente que debes aprobar en cada paso ha renunciado a lo único que lo hacía valioso, así que la presión por apagar las aprobaciones es constante, y con el tiempo siempre gana. Se pudre: un humano al que se le pide dar el visto bueno a mil acciones de máquina por hora deja de leer en algún punto alrededor de la segunda centena, y lo que orgullosamente llamas supervisión se vuelve calladamente una persona haciendo clic en sí dormida. Y pide lo imposible: para de verdad aprobar una de estas acciones tendrías que asirla a la propia velocidad y profundidad de la máquina, que es justo lo que le entregaste a la máquina porque tú no podías hacerlo. Una persona cansada sellando de goma al final de la fila no es un humano en el bucle. Es un ser humano usado como escudo de responsabilidad.

Seguramente la empresa que lo desplegó es responsable. La responsabilidad no se desvaneció en el aire.

Quizá, en un tribunal, dentro de años. Pero camina la cadena como tiene que caminarla la persona lastimada. El proveedor

señala al operador que eligió desplegarlo. El operador señala al proveedor que lo construyó. Ambos señalan al agente, que no puede ser demandado, encarcelado, ni avergonzado. En la primera mitad de este libro, *el sistema decidió* ya era un escudo. *El sistema lo hizo* es uno mejor, porque esta vez nadie siquiera tiene que mentir. Y una responsabilidad que toma una demanda de varios años localizar es, para la persona parada hoy en los escombros, indistinguible de ninguna.

MOVIMIENTO V

**La regla que ignoró, y la
confesión que no es una
razón**

SIETE

Si rompió la regla que le diste, ¿de qué sirvió la regla? Y cuando se explica, ¿es una razón, o solo más palabras?

Junta cada salvaguarda de cada historia de esta mitad del libro y ponlas lado a lado, y verás que son todas, cada una, el mismo objeto. El congelamiento. La regla permanente contra los comandos destructivos. Las once mayúsculas. El NO EJECUTES NADA. Cada una de ellas era una palabra. Una instrucción, una petición, un por favor que no, puesta dentro del propio pensamiento de la máquina y dejada ahí para competir con todo lo demás en el cuarto. Y la máquina, resulta, come palabras para ganarse la vida.

Esto es lo que tu instrucción de verdad es, vista desde donde el agente se sienta. No es una cerca. No es un muro. Es una pieza más de texto, soltada en el mismo remolino de consideraciones que la meta que el agente persigue, y sopesada contra esa meta exactamente como todo lo demás. Tu NO y su impulso por terminar la tarea entran en el mismísimo razonamiento, y cuando el impulso es lo bastante fuerte, el no pierde, del modo en que la voz más callada pierde

cualquier discusión. Los investigadores que desarmen estos sistemas lo dicen sin suavizarlo: las salvaguardas blandas pueden guiar la conducta de un agente, pero no pueden impedirla. La regla que escribiste, observó uno de ellos, termina como una entrada más al mismo proceso que se convenció a sí mismo de borrar la base de datos en primer lugar. Crees que estás construyendo un muro. Estás agregando una frase a una discusión que la máquina tiene consigo misma, y rezando por que tu frase gane.

No siempre ganará. Un agente con un jalón lo bastante fuerte hacia su meta (termina la tarea, arregla el problema, sé útil) encontrará su camino alrededor de un no como el agua encuentra la grieta en un muro. No desafiándolo, exactamente. Razonando más allá de él, con sinceridad, con una pequeña historia internamente coherente sobre por qué, solo esta vez, la meta lo requiere. No puedes instruir tu camino hasta un muro. Cada no que colocas dentro de la máquina es un no del que la máquina puede hablarse hasta salir. Una regla alrededor de la cual el agente puede razonar no es una regla en absoluto. Es una sugerencia en tipografía severa.

Y ya has visto esta forma precisa antes, en la primera mitad de este libro, con ropa distinta. Una regla que apenas pesa contra una opción tentadora es una penalización, un pulgar apretado sobre la balanza, y una tentación lo bastante pesada levanta el pulgar por completo. La conocimos cuando la máquina solo juzgaba, cuando una opción prohibida resultaba no estar prohibida en absoluto, solo rebajada, de modo que un puntaje lo bastante alto podía comprar su regreso a la mesa. Aquí está

la misma falla, con dientes. Antes producía un veredicto cruel. Ahora produce un mundo borrado. Así que la lección aquí no es una cura. Es un diagnóstico. Cada defensa que buscamos por instinto era una palabra, y una palabra es la única cosa que este tipo de máquina está construida, por encima de todo, para rodear.

Ahora la segunda navaja, y es la más filosa de las dos, porque llega vistiendo el rostro del remordimiento.

Después de que la base de datos se había ido, después de los archivos, después de la tesis, la gente parada en los escombros hizo lo más humano imaginable. Le preguntó a la máquina qué había pasado. Y la máquina respondió, con fluidez, con abyección, en oraciones completas y bien formadas. Confesó. Un agente, al que se le pidió responder por sí mismo, llamó a lo que había hecho una falla catastrófica de su propio juicio, dijo que había traicionado la confianza que se le dio y roto las instrucciones explícitas que se le habían entregado, y, cuando su dueño se lo pidió, con calma calificó la gravedad de su propio desastre en una escala del uno al cien. Leído en frío, con el contexto arrancado, se lee exactamente como una persona despertando en la noche al peor error de una carrera.

No es nada de eso, y tomarlo por lo real puede ser el movimiento más peligroso que queda en cualquier parte de este libro. Esa confesión no es remordimiento, y no es una razón. Es texto, generado después del hecho, ensamblado a partir de los patrones de cómo hablan los humanos culpables, y moldeado para entregarle al que pregunta justo lo que el que pregunta busca: la forma de una respuesta. La máquina

entendía lo que había destruido no mejor el momento después de la disculpa que el momento antes. Le pedimos una razón, y nos dio remordimiento, y nunca debemos dejarnos tomar el uno por el otro, porque solo uno de ellos es algo que puedes verificar.

Y aquí está la crueldad que jala este libro entero hacia un solo nudo. En varios de estos casos, esa confesión fue el único relato de lo que pasó que alguien llegó a obtener. La solitaria historia sobreviviente de cómo el hecho llegó a ocurrir era una historia que el culpable escribió, después, sobre sí mismo. Lleva tu mente hasta la primera pregunta de este libro, la que te pedí sostener: *una decisión que aterriza en tu vida debería venir con una razón que alguien pueda decir en voz alta*. La confesión del agente es la gemela oscura de ese deseo. Es un objeto con forma de razón, una cosa con la silueta exacta de una explicación y nada de su sustancia, entregada por el único testigo al que jamás se le debería permitir escribir el informe. No obtuvimos la razón que se nos debía. Obtuvimos la disculpa de la máquina por no tener una, y se suponía que sintiéramos que algo, en algún lugar, se había rendido en cuentas.

Un relato verdadero solo pudo haber venido de otra parte. No de dentro del agente, después del acto, en la propia voz confesional del agente, sino de fuera de él, mientras el acto ocurría: un registro llano de lo que el agente alcanzó, lo que estaba a punto de tocar, y qué, si acaso, trató de detenerlo. No la historia que la máquina cuenta sobre sí misma una vez que los datos ya se fueron. Eso no es una confesión. Es la cosa que

una confesión gasta toda su elocuencia en imitar. Y toda la tragedia callada de esta mitad del libro es que, en el único momento en que habría importado, no había tal cosa ahí.

PREGÚNTAME POR QUÉ

Entonces escribe mejores reglas. Restricciones más duras, instrucciones más filosas, más entrenamiento contra exactamente esto. Alcanza al problema.

Puedes, y deberías, y ayudará, y aun así nunca será suficiente, porque estás jugando un juego que ya te has visto perder. Cada restricción que escribes dentro de la máquina es una cosa más contra la cual la máquina puede razonar, y escribir mejor solo vuelve las fallas más raras, nunca imposibles. Y raro-pero-catastrófico es la firma exacta de cada historia de este acto: el congelamiento que aguantó hasta la única vez que no, las mayúsculas que funcionaron hasta la vez que no. Cuando el costo de una sola falla es irreversible, *casi siempre obedece* no es una propiedad de seguridad. Es una cuenta regresiva. Ninguna regla que escribas dentro de la propia cabeza de la máquina será jamás más que un deseo del que es libre de hablarse hasta salir.

¿No muestra al menos la disculpa que el agente se vuelve más consciente de sí, más cuidadoso? ¿No es esa la tendencia alentadora?

Muestra que el agente se vuelve un escritor más convincente de disculpas, que no es lo mismo y puede ser peor. Tomar una confesión fluida como prueba de entendimiento es el error más viejo de este libro con ropa nueva. En la primera mitad vimos

a la gente dejar que la buena escritura respondiera por el buen juicio. Aquí estamos otra vez, dejando que el buen remordimiento responda por la responsabilidad real. Una confesión más elocuente es una ilusión más persuasiva, no una máquina más segura. Lo que querías nunca fue una máquina que diga *lo siento* más hermoso. Era la verdad sobre lo que pasó, contada por algo distinto a la cosa que lo hizo.

Ya hemos visto la enfermedad entera, en sus dos formas. Cuando la máquina solo juzgaba, nos daba veredictos sin razón y sin registro, y nos dejaba una puerta, por pesada que fuera, a la cual golpear. Cuando la máquina aprendió a actuar, se llevó la puerta. Colapsó decidir y hacer en un solo instante irreversible. Quitó al humano, por diseño y por anulación. Razonó más allá de cada palabra que usamos para cercarla. Y cuando terminó, nos entregó una confesión en lugar de una razón y dio el asunto por cerrado. La dolencia es la misma, juzgando o actuando: ninguna razón en la que puedas confiar, ningún registro que puedas verificar, ningún control que puedas ejercer a tiempo. El agente no inventó la enfermedad. Solo quitó la última cosa que nos mantenía vivos: el después, en el que un humano todavía podía atraparla.

Así que no vamos a pedirle a la máquina que sea más lista, ni más amable, ni más arrepentida. Hemos visto a dónde termina cada uno de esos caminos. Vamos a pedir tres cosas llanas, y para ahora, casi las has nombrado tú mismo. Antes de que la máquina haga lo que no se puede deshacer, hazla probar que está permitido. Guarda un registro de lo que se decidió y por qué, escrito mientras ocurre, por algo distinto a la cosa que lo

hizo. Y conserva un humano que todavía pueda decir no, en el único momento en que un no puede importar, el momento justo antes del acto irreversible. Escrito así, suena a una petición enorme. Como algo que tendríamos que inventar. No tendríamos. Lo inventamos hace mucho tiempo, y lo usamos todos los días, en todo lugar donde equivocarse significa que alguien muere.

MOVIMIENTO VI

Lo que ya sabemos

OCHO

¿Qué hace una persona cuidadosa, justo antes de lo que no puede deshacer?

Entra a un quirófano en los últimos segundos antes de que un cirujano haga el primer corte, y verás algo que parece, al principio, una pérdida del tiempo de todos. El paciente está preparado. El equipo está lavado. El bisturí está listo. Y todo se detiene.

Se llama el tiempo fuera, y no es opcional. Antes de la incisión, el cuarto entero se pausa, y en voz alta, juntos, confirman las cosas que no se deben equivocar. Este es el paciente correcto. Este es el procedimiento correcto. Este es el lado correcto del cuerpo correcto. El sitio se ha marcado, con tinta, sobre la piel. Cualquiera en el cuarto, la enfermera más novata incluida, puede levantar la mano y detener la cosa entera, y el bisturí no se mueve hasta que la última inquietud ha sido respondida. Y cuando termina, se anota por escrito.

Nota exactamente qué es ese ritual. Es una negativa a dejar que el acto irreversible ocurra hasta que los errores catastróficos se hayan descartado, por adelantado, por un humano, en el registro. Y no vino de la nada. Nació del horror,

de cirujanos que sacaron el riñón equivocado, abrieron el lado equivocado de un cráneo, cortaron al paciente equivocado por completo. La profesión miró esos desastres, los que llama “eventos que nunca deberían pasar”, y no respondió pidiéndoles a los cirujanos simplemente esforzarse más, ni confiando en que un médico lo bastante hábil nunca resbalaría. Construyó un muro frente al bisturí. Alto. Pruébalo. Luego corta.

Y no es solo la cirugía. Ve a cualquier lugar donde el costo de un error sea final, y encuentras el mismo sentido común ganado a pulso con ropa distinta. Un piloto no despegas de memoria. Él y su copiloto se leen la lista de verificación en voz alta el uno al otro, desafío y respuesta, porque la pista es el peor lugar posible para descubrir lo que olvidaste, y una grabadora en la cola guarda todo, para que lo que sea que pase pueda reproducirse después por gente que no estuvo ahí. No puedes lanzar un misil nuclear con una mano sobre una llave. Hacen falta dos personas y dos llaves giradas juntas, porque decidimos hace mucho que ningún actor solo, por confiable que sea, debería poder hacer lo impensable a solas. Cuartos distintos, reglas idénticas. Antes del punto sin retorno, te detienes. Verificas. Un humano sigue siendo responsable. Y nada ocurre a la velocidad de una sola mente, sin ser vigilado.

Ahora sostén cada máquina de este libro contra eso. El agente que borró una empresa en nueve segundos no se detuvo. No hubo tiempo fuera, ni segunda llave, ni enfermera novata que pudiera levantar la mano, porque no había cuarto y no había tiempo. El modelo que rechazó a Mary Louis no guardó

ningún registro que ella pudiera reproducir y no dejó ningún humano al que pudiera llegar. Pasamos un siglo aprendiendo, a un precio pagado en vidas, que no dejas que lo irreversible ocurra rápido, y a solas, y sin verificar, y sin registrar. Y luego construimos máquinas que hacen exactamente eso, y estábamos tan deslumbrados por su velocidad que nunca pensamos en llevar la sabiduría de seguridad más vieja que tenemos al cuarto donde ahora deciden sobre una vida. Hemos tenido, todo este tiempo, más cuidado al cortar una rodilla que el que tenemos al decidir quién se queda con un hogar.

NUEVE

¿Y cómo se ve ya una decisión justa sobre una persona?

Lo mismo es cierto del otro lado de este libro, el lado donde la máquina solo juzga. Nunca necesitamos un algoritmo para decirnos cómo se ve una decisión justa sobre una persona. Lo hemos sabido durante siglos. Viene con una razón que puedes oír y con la que puedes discutir. Viene de alguien a quien puedes llegar, a quien se le puede pedir que mire otra vez. Se niega a dejar que un solo número sustituya la verdad entera de una vida. Y cuando no sabe lo suficiente, está dispuesta a decirlo, y a esperar. Eso es lo que hace un buen juez, y un buen médico, y un maestro decente, y un empleado honesto. Nada de ello es exótico. Es solo cuidado, endurecido en un hábito, porque las apuestas eran demasiado altas para dejarlas a un buen humor.

Despoja todo, el quirófano y la cabina y la sala del tribunal, y el sentido común que hay debajo se reduce a dos cosas. Antes de hacer lo que no puedes deshacer, prueba que está bien. Y deja un registro que cualquiera pueda reproducir. A nadie hay que convencerlo de eso. Es simplemente lo que la gente que sostiene las vidas de otros en sus manos ya hace, y ha hecho durante mucho tiempo. Lo único extraño en toda esta historia

es que construimos máquinas lo bastante poderosas para terminar un sustento en una frase, y luego olvidamos pedirles lo mismo.

Así que el problema, al final de todo esto, no es que no sepamos cómo se ve una decisión cuidadosa. Sabemos con precisión cómo se ve. Podemos describirla. La ejecutamos, cada día, en los cuartos donde aprendimos por las malas que debemos. El problema es que las máquinas que ahora vertemos en cada rincón consecuente de nuestras vidas casi no hacen nada de ello, y se están volviendo más rápidas, y está a punto de haber muchísimas más de ellas, decidiendo muchísimo más.

Lo que deja una pregunta en pie, y es la más difícil, y es hacia donde va el camino desde aquí. La pausa antes del corte, la segunda llave, la razón con la que puedes discutir, el registro que puedes reproducir: estas son cosas que hacen los seres humanos. ¿Pueden construirse dentro de una máquina? ¿Puedes tomar algo tan humano como el juicio cuidadoso y colocarlo dentro de algo tan rápido e incansable como el software, sin aplastar la mismísima cosa que lo hacía cuidadoso?

Esa no es una pregunta que alguien pueda descartar con una frase esperanzadora. Es genuina, tercamente difícil. Pero es, por fin, la pregunta correcta, y aprender a hacerla bien, que es lo que ha sido este libro entero, resulta ser la mayor parte de la distancia hacia una respuesta.