Más control, más caos

Límites externos para sistemas que no se detienen solos

Ulises González

2025

© 2025 Ulises González. Todos los derechos reservados.

Primera edición: 2025

Índice general

Capítulo 1

IA y los límites humanos

Una organización implementa un modelo predictivo para optimizar decisiones de inventario. El modelo funciona exactamente según especificaciones: procesa datos históricos, identifica patrones, genera recomendaciones de reabastecimiento con precisión superior a la del equipo humano anterior. Los indicadores mejoran en los primeros trimestres. El costo de inventario baja. Los quiebres de stock se reducen. El éxito se celebra internamente y se presenta al directorio como validación de la estrategia de automatización. Nadie nota que el modelo está optimizando para condiciones de mercado que ya no existen porque los datos que lo entrenan tienen un rezago estructural que nadie definió como problema. Cuando el mercado cambia de manera que los patrones históricos dejan de predecir el futuro, el modelo sigue recomendando con la misma confianza de siempre. Este fenómeno se conoce en la literatura como *concept drift*: la degradación silenciosa de modelos cuando cambian las condiciones que los entrenaron. Bayram et al. (2022) documentan cómo los modelos continúan produciendo outputs con alta confianza aparente mientras su validez predictiva se deteriora—sin mecanismo interno para detectar su propia obsolescencia. Los indicadores tardan meses en reflejar el deterioro porque el modelo no tiene forma de saber que está equivocado. Para cuando el problema es visible, el inventario acumulado representa pérdidas que superan varios años de los ahorros que el modelo generó.

Otra organización despliega un sistema de scoring para priorizar oportunidades comerciales. El sistema aprende de decisiones pasadas del equipo de ventas y replica sus patrones a escala. Lo que nadie explicitó es que las decisiones pasadas contenían sesgos que el equipo desconocía o que consideraba irrelevantes. El sistema ahora aplica esos sesgos de manera consistente y documentada sobre un volumen de decisiones que ningún humano podría revisar individualmente. El resultado agregado es una concentración de cartera que el equipo de riesgos detecta tarde, cuando ya representa exposición significativa. El sistema no introdujo el sesgo. Lo amplificó hasta hacerlo visible de manera que antes no era posible. Mehrabi et al. (2021) documentan cómo los algoritmos de ML no solo replican sesgos presentes en datos de entrenamiento sino que los amplifican sistemáticamente—un fenómeno que O'Neil (2016) denomina la "weaponización" del sesgo a escala computacional.

Una tercera organización implementa dashboards automatizados que sintetizan información de múltiples fuentes y presentan al comité ejecutivo una vista consolidada del negocio. Los ejecutivos reciben reportes más frecuentes, más detallados, más visualmente atractivos. La sensación de control aumenta porque hay más información disponible más rápido. Parasuraman y Manzey (2010) denominan este fenómeno *automation bias*: la tendencia cognitiva a aceptar outputs automatizados como heurístico de reemplazo para el procesamiento crítico de información. Los dashboards más sofisticados pueden intensificar este sesgo al crear apariencia de exhaustividad que reduce el escrutinio. Lo que no aumenta es la capacidad de evaluar si esa información es relevante para las decisiones que importan. Los dashboards muestran lo que el sistema fue diseñado para mostrar, no necesariamente lo que el comité necesita ver. La proliferación de métricas crea la ilusión de comprensión exhaustiva mientras oscurece las preguntas que nadie está haciendo porque no aparecen en ningún indicador automatizado.

Una institución financiera regional automatizó su proceso de evaluación crediticia replicando los criterios que su equipo comercial había usado durante décadas. El modelo era técnicamente impecable: procesaba solicitudes en minutos, reducía costos operativos, eliminaba variabilidad entre analistas. Lo que el modelo también replicaba, a escala y velocidad que ningún equipo humano había alcanzado, eran sesgos de concentración geográfica y sectorial que el equipo original había desa-

rrollado orgánicamente sin documentarlos como criterio explícito. Cuando el ciclo económico cambió y los sectores sobreexpuestos entraron en estrés, la cartera deterioró a una velocidad que el área de riesgos no había modelado porque nadie sabía que la concentración existía en esa magnitud. El modelo no había creado el sesgo. Lo había escalado hasta hacerlo sistémicamente relevante.

Estos patrones no son fallas de la tecnología. Son fallas de decisiones humanas que la tecnología ejecutó fielmente y a escala. El modelo de inventario no decidió ignorar cambios de mercado; nadie le indicó que los considerara. El sistema de scoring no decidió concentrar riesgo; replicó lo que los humanos habían hecho antes de manera menos visible. Los dashboards no decidieron ocultar información crítica; mostraron exactamente lo que se les pidió mostrar. La inteligencia artificial no es inherentemente peligrosa ni inherentemente beneficiosa. No tiene agencia propia para hacer daño ni para generar valor. No toma decisiones en ningún sentido significativo de la palabra. Ejecuta instrucciones codificadas por humanos sobre datos seleccionados por humanos para optimizar objetivos definidos por humanos. Cuando los resultados son problemáticos, la causa no está en la tecnología sino en las decisiones humanas que la tecnología amplificó.

Esta distinción es crítica porque cambia completamente dónde buscar soluciones. Si el problema fuera la IA misma, la respuesta sería limitar la IA, regularla, frenarla, quizás prohibirla en ciertos contextos. Pero si el problema son decisiones humanas mal definidas que la IA escala eficientemente, la respuesta es mejorar las decisiones humanas antes de automatizarlas. La tecnología es neutral respecto a la calidad de lo que amplifica. Amplifica igualmente bien decisiones correctas y decisiones problemáticas. La diferencia en resultados depende enteramente de lo que se le pide amplificar.

Los ejecutivos que implementaron los sistemas descritos en la sección anterior no eran irresponsables ni incompetentes. Actuaron con la información disponible, siguieron procesos razonables, tomaron decisiones que parecían correctas en su momento. El problema no fue falta de diligencia individual. Fue que los límites de lo que el sistema automatizado podía y no podía hacer nunca fueron explicitados de manera que permitiera anticipar los modos de falla que eventualmente ocurrieron. Nadie definió bajo qué condiciones el modelo de inventario debería dejar de ser con-

fiable. Nadie especificó qué sesgos del equipo de ventas no debían replicarse. Nadie determinó qué preguntas críticas los dashboards debían responder aunque nadie las hubiera formulado explícitamente.

La ausencia de estos límites no fue negligencia. Fue el estado normal de organizaciones que no habían necesitado explicitarlos antes porque la escala humana de operación hacía que los errores fueran detectables y corregibles antes de acumularse. La IA eliminó esa protección implícita al permitir que las decisiones se ejecutaran a una escala donde la detección humana ya no podía operar. La razón estructural por la cual la IA expone límites humanos que antes permanecían ocultos tiene que ver con una asimetría fundamental: la capacidad técnica de procesar y ejecutar crece exponencialmente mientras que la capacidad humana de establecer criterios, evaluar consecuencias y definir límites permanece constante.

Los humanos tienen atención finita. Pueden monitorear un número limitado de variables simultáneamente. Pueden evaluar un número limitado de decisiones por unidad de tiempo. Pueden anticipar consecuencias de segundo y tercer orden solo hasta cierto punto de complejidad. Estas limitaciones no son defectos que la tecnología vaya a corregir. Son características estructurales de la cognición humana que ninguna herramienta elimina. Endsley (2017) documenta la "paradoja de la supervisión": conforme aumenta la automatización, disminuye la capacidad humana de mantener awareness situacional suficiente para intervenir efectivamente cuando el sistema falla. Lo que la IA hace es permitir que se tomen y ejecuten decisiones a una escala que excede dramáticamente la capacidad humana de supervisión significativa.

Cuando un equipo humano tomaba decisiones de inventario manualmente, cada decisión pasaba por un proceso cognitivo que, aunque imperfecto, incluía cierta evaluación contextual. El analista que recomendaba una compra grande podía notar que algo en el mercado había cambiado aunque no supiera exactamente qué. La intuición desarrollada por años de experiencia funcionaba como un sistema de alerta temprana impreciso pero real. Polanyi (1966) denomina este tipo de expertise "conocimiento tácito"—el saber-cómo que se adquiere solo mediante práctica sostenida. Rinta-Kahila et al. (2018) documentan cómo la automatización erosiona este conocimiento al eliminar las oportunidades de ejercicio que lo mantienen: el deski-

lling es efecto latente que solo se hace visible cuando el sistema automatizado falla. Cuando esas mismas decisiones las toma un modelo automatizado, la evaluación contextual desaparece porque el modelo no tiene intuición ni capacidad de notar lo que no fue programado para notar. La decisión se ejecuta sin el filtro humano que antes operaba de manera invisible.

La IA no decide mal. Ejecuta decisiones mal definidas de manera eficiente. La distinción es crucial. Una decisión mal definida tomada por un humano tiene alcance limitado y es corregible cuando las consecuencias se hacen visibles. La misma decisión mal definida ejecutada por un sistema automatizado tiene alcance potencialmente ilimitado y puede acumular consecuencias durante mucho tiempo antes de que sean detectables. El problema no es la velocidad de ejecución ni la escala de operación. Es que la velocidad y la escala magnifican las consecuencias de definiciones incompletas que antes tenían impacto manejable. Bainbridge (1983) articuló esta dinámica como las "ironías de la automatización": automatizar lo fácil deja a los humanos las tareas difíciles—intervenir cuando el sistema falla—pero simultáneamente los despoja de la práctica necesaria para hacerlo competentemente. La IA contemporánea intensifica estas ironías a escala sin precedentes.

Los límites humanos siempre existieron. La IA no los creó. Los hizo visibles al eliminar los mecanismos implícitos que antes los compensaban parcialmente. El riesgo específico de introducir IA en sistemas organizacionales que no tienen límites explícitos no es el riesgo genérico de la tecnología ni el riesgo abstracto de la automatización. Es la aceleración de dinámicas que este libro ha descrito desde el primer capítulo.

El loop de amplificación que comienza con energía organizacional y se autorefuerza hasta encontrar un límite externo opera ahora a velocidad aumentada. Una iniciativa que antes tardaba meses en acumular momentum suficiente para ser indetenible puede ahora acumular ese momentum en semanas porque la IA acelera cada paso del proceso. Los reportes se generan más rápido, las métricas se actualizan en tiempo real, las proyecciones se refinan continuamente. Todo el aparato de justificación que sostiene el momentum se vuelve más eficiente sin que la capacidad de cuestionar ese momentum aumente proporcionalmente.

La opacidad decisional crece porque las decisiones que antes eran visibles y

cuestionables ahora están embebidas en modelos que pocos entienden y nadie revisa sistemáticamente. Un comité ejecutivo puede cuestionar la recomendación de un director que presenta un análisis en una reunión. Es mucho más difícil cuestionar la salida de un sistema automatizado que presenta esa misma recomendación respaldada por miles de data points procesados de maneras que nadie en la sala puede explicar completamente. La autoridad epistémica se traslada del juicio humano visible al algoritmo invisible sin que nadie haya decidido explícitamente que eso era deseable. Burrell (2016) identifica tres fuentes de opacidad algorítmica: complejidad técnica, secreto corporativo, y la naturaleza inherentemente no-intuitiva de cómo los modelos de ML procesan información.

La reversibilidad disminuye porque las consecuencias de decisiones automatizadas se acumulan más rápido de lo que pueden corregirse. Cuando un error humano produce consecuencias visibles, usualmente hay tiempo para detectar el problema y corregir el curso antes de que el daño sea irreversible. Cuando un error de configuración en un sistema automatizado produce consecuencias, esas consecuencias pueden acumularse durante el tiempo que tarda alguien en notar que algo anda mal, y para entonces el costo de reversión puede exceder el costo de las consecuencias mismas. Perrow (1984) describe cómo los sistemas altamente automatizados y acoplados pueden acumular fallos más rápido de lo que los operadores humanos pueden detectarlos—una dinámica que denomina "accidentes normales". La automatización por IA intensifica este acoplamiento mientras reduce las señales visibles de deterioro.

La IA no crea estos riesgos de la nada. Amplifica riesgos que ya existían en la estructura organizacional pero que operaban a una escala donde eran manejables. El ejecutivo que antes podía confiar en que los errores serían detectables a tiempo ya no puede confiar en eso cuando la velocidad de ejecución excede la velocidad de detección humana. La protección frente a la amplificación de límites humanos por IA no consiste en limitar la IA sino en explicitar los límites humanos antes de que la IA los encuentre por ensayo y error costoso. Esto conecta directamente con todo lo que este libro ha establecido sobre decisiones, aprendizaje y mecanismos de límite.

El Decisión Readiness Gate opera como filtro previo a cualquier automatización significativa. Una iniciativa que propone implementar IA para optimizar algún

proceso organizacional debe pasar por el gate con criterios específicos sobre qué límites humanos están en juego y cómo se manejarán. El gate no evalúa si la IA es técnicamente viable ni si los beneficios proyectados son atractivos. Evalúa si las decisiones que la IA va a ejecutar a escala están suficientemente bien definidas como para que la amplificación produzca resultados deseables en lugar de amplificar errores latentes.

El Capítulo 3 describió cómo la IA acelera el Coding Trance. Lo que sigue traduce esa observación en criterios operativos. Las iniciativas que involucran delegación algorítmica de decisiones requieren escrutinio específico que las iniciativas tradicionales no requieren. No porque la IA sea inherentemente más riesgosa, sino porque amplifica más rápido, falla de maneras menos visibles, y es más difícil de revertir una vez desplegada. Lo que sigue no son criterios exhaustivos. Son preguntas mínimas que cualquier iniciativa de este tipo debería poder responder antes de recibir autorización de ejecución. La incapacidad de responderlas no es señal de que la iniciativa sea mala; es señal de que no está lista.

Primera pregunta: qué decisiones humanas replica. Un modelo que optimiza sin claridad sobre qué decisión humana sustituye no puede ser evaluado. La pregunta no es qué hace el modelo técnicamente; es qué juicio humano deja de ejercerse porque el modelo existe. La señal de no-readiness es cuando el equipo describe el modelo en términos de arquitectura técnica pero no puede articular qué decisión humana específica el modelo está tomando. El modelo predice X no es respuesta. El modelo decide si Y recibe Z es respuesta.

Segunda pregunta: qué sesgos hereda. Los modelos aprenden de datos históricos. Los datos históricos contienen decisiones humanas. Las decisiones humanas contienen sesgos. El modelo no elimina sesgos; los escala. Un sesgo que afectaba cien decisiones manuales por mes ahora afecta diez mil decisiones automatizadas por hora. La señal de no-readiness es cuando el equipo asume que el modelo es objetivo porque es matemático, o que el sesgo se resuelve con más datos. No hay análisis de qué patrones históricos problemáticos el modelo podría estar replicando.

Tercera pregunta: bajo qué condiciones el modelo deja de ser confiable. Todo modelo tiene supuestos sobre el mundo en el que opera. Cuando el mundo cambia, el modelo sigue produciendo outputs con la misma confianza pero menor validez.

La pandemia invalidó modelos de demanda entrenados en datos pre-pandemia. El modelo no sabe que está equivocado; sigue prediciendo con precisión aparente. La señal de no-readiness es cuando no hay definición de qué cambios en el contexto invalidarían los supuestos del modelo. No hay mecanismo de monitoreo que detecte drift entre las condiciones de entrenamiento y las condiciones actuales.

Cuarta pregunta: quién tiene autoridad de apagado. Los sistemas automatizados tienden a permanecer encendidos por default. El costo de apagarlos es visible e inmediato. El costo de mantenerlos encendidos cuando no deberían es difuso y diferido. Sin autoridad explícita de apagado, el modelo sigue operando hasta que produce daño visible. La señal de no-readiness es cuando la autoridad de apagado es vaga o inexistente. No hay criterio definido de qué constituye divergencia suficiente para activar revisión. No hay proceso documentado de qué pasa después del apagado.

Quinta pregunta: puede el equipo explicar el impacto en lenguaje no técnico. Si el equipo no puede explicar cómo una decisión del modelo afecta a una persona específica, no puede evaluar si ese impacto es aceptable. El modelo optimiza para X no es explicación de impacto. Si el modelo te clasifica como Y, entonces Z te ocurre es explicación de impacto. La señal de no-readiness es cuando las explicaciones son exclusivamente técnicas. El equipo no ha mapeado la cadena causal desde output del modelo hasta consecuencia para el afectado.

Estas preguntas no son lista de verificación de cumplimiento. Son filtro mínimo de readiness. Una iniciativa que no puede responderlas no está lista para aprobación, independientemente de la presión por ejecutar. El DRG debe incluir estos criterios como parte de su evaluación de cualquier iniciativa que involucre delegación algorítmica. No como sección separada del formulario, sino como profundización del escrutinio estándar. La IA no es excepción al proceso; es caso que requiere más proceso.

El Apéndice A incluye estos criterios integrados en la categoría de iniciativas de datos e IA. Esos criterios traducen los principios abstractos de este capítulo en verificaciones concretas que el gate puede aplicar: existencia de gobernanza de datos documentada, definición de métricas de sesgo aceptable, planes de monitoreo post-despliegue, criterios de reversión si los resultados divergen de lo esperado. La

diferencia entre una iniciativa de IA que amplifica fortalezas y una que amplifica debilidades frecuentemente se reduce a si estos criterios se verificaron antes de autorizar inversión de escala.

El aprendizaje procedural que el capítulo anterior describió es condición necesaria para que la IA produzca valor sostenible. Un sistema automatizado que replica decisiones humanas pasadas solo es tan bueno como esas decisiones. Si las decisiones pasadas contenían errores que la organización no ha codificado como reglas a evitar, el sistema automatizado replicará esos errores a escala. Si el aprendizaje de fracasos anteriores quedó en memorias individuales en lugar de criterios codificados, el sistema automatizado no tendrá acceso a ese aprendizaje y repetirá los mismos patrones que causaron problemas antes.

El veredicto RECHAZO del DRG adquiere importancia adicional cuando la iniciativa bajo evaluación involucra IA. Detener una automatización mal diseñada antes de que entre en producción evita no solo las consecuencias directas del error sino la amplificación de esas consecuencias que la IA habría producido. El costo de un RECHAZO temprano es trivial comparado con el costo de descubrir tarde que un sistema automatizado estuvo amplificando decisiones problemáticas durante meses o años.

La IA no sustituye el juicio humano. Hace visible dónde el juicio humano nunca estuvo, dónde las decisiones se tomaban por inercia o precedente sin que nadie explicitara los criterios que supuestamente las gobernaban. Cuando un sistema automatizado produce resultados problemáticos, casi siempre revela decisiones que los humanos tomaban mal de manera menos visible. La IA no creó el problema; lo iluminó a una escala donde ya no puede ignorarse.

Hay una implicación adicional de está visibilidad que merece atención explícita. Los sesgos humanos distribuidos entre múltiples decisores eran difíciles de detectar porque cada instancia era pequeña y la suma agregada no era visible para nadie. Un analista que favorece ciertos tipos de clientes, otro que evita ciertos sectores, un tercero que pondera riesgo de manera conservadora: la cartera resultante refleja la suma de estos sesgos individuales pero nadie puede señalar un punto específico donde el sesgo se introdujo. La IA hace estos sesgos consistentes y documentados. Un modelo que replica el sesgo agregado del equipo humano produce una cartera

donde el sesgo es medible, atribuible, auditable.

Esta visibilidad tiene dos caras. Para organizaciones sin mecanismos de límite, significa que los errores latentes ahora producen evidencia que puede usarse en contra. Para organizaciones con DRG y aprendizaje codificado, significa que los sesgos antes invisibles ahora pueden identificarse, discutirse y corregirse antes de que produzcan daño material. La misma tecnología que amplifica el riesgo para unos reduce el riesgo para otros, dependiendo enteramente de si la organización tiene la arquitectura institucional para procesar lo que la IA hace visible.

La IA no elimina el error humano. Elimina la excusa de no haberlo visto venir. Para la organización que tiene límites externos y criterio codificado, eso es una oportunidad. Para la que no los tiene, es una exposición que antes no existía.

La organización que ha instituido el DRG como límite externo, que ha codificado su aprendizaje en criterios procedurales, que sabe producir veredictos negativos antes de que sea demasiado tarde, puede integrar IA de manera que amplifique sus fortalezas en lugar de sus debilidades. La organización que carece de estos mecanismos encontrará que la IA amplifica exactamente lo que menos quiere amplificar: las decisiones mal definidas, los sesgos no reconocidos, los límites humanos que nadie explicitó porque nadie pensó que sería necesario.

Poner límites humanos explícitos no frena la IA. Evita que la IA acelere lo que nunca tendría que haber existido.