STRATEX

La estrategia que crea valor

El costo invisible de la IA:

lo que las empresas arriesgan si delegan el juicio crítico a la IA

Cómo gobernar la rendición cognitiva en proyectos de IA

El costo invisible de la IA:

El riesgo detrás de una adopción sin control

por Stratex

La idea en síntesis

La IA está entrando en los procesos de decisión más rápido de lo que las empresas la pueden gobernar. La evidencia es contundente: cuando el modelo se equivoca, los usuarios lo siguen el 80% de las veces, y su confianza aumenta.	El fenómeno no es delegación cognitiva —usar la IA como herramienta bajo control humano— sino rendición cognitiva: el operador valida la respuesta del modelo sin chequeos y la hace propia.	Los casos de negocio de IA asumen que el operador humano corregirá el error del modelo. No lo hace: lo firma. Y al validarlo, el fallo técnico se convierte en decisión organizacional.	Mitigarlo exige tres decisiones: gobernar la IA como activo, no como software; clasificar cada caso de uso por zona de riesgo (A/B/C) antes de aprobarlo; y medir corrección humana, no adopción.

Resumen ejecutivo

A medida que la IA entra en los procesos de decisión empresarial, aparece una forma elegante de equivocarse que las organizaciones aún no han aprendido a detectar. No es el error técnico del modelo, ni el error obvio que un usuario atento detecta y corrige. Es algo más sutil: el momento en que un ejecutivo lee la respuesta de un asistente de IA, la valida, y se la apropia. No verificó. No contrastó. No la sometió a la prueba que habría aplicado a una recomendación humana. Y, sin embargo, quedo convencido que su decisión era correcta.

Un trabajo reciente de Steven Shaw y Gideon Nave, de Wharton, pone número a este fenónmeno.[1] En tres experimentos con 1.372 participantes y casi diez mil ensayos, los autores midieron qué ocurre cuando un asistente de inteligencia artificial entrega respuestas a problemas de razonamiento, manipulando deliberadamente —mediante instrucciones ocultas— si la respuesta es correcta o errónea. El hallazgo central es incómodo. Cuando los participantes consultaban al asistente y éste se equivocaba, lo seguían en cuatro de cada cinco ocasiones. La precisión humana, que en condiciones normales rondaba el 46%, caía al

31% en presencia de una IA defectuosa, por debajo de quienes resolvían el mismo problema sin asistencia alguna. La confianza, en cambio, subía once puntos. Tener acceso a un asistente hacía que la gente se sintiera más segura, incluso cuando este entregara respuesta erróneas.

Los autores llaman a este fenómeno *rendición cognitiva*, y lo distinguen con cuidado de algo más antiguo y más benigno: la delegación cognitiva (*cognitive offloading*), un concepto que la psicología cognitiva conoce desde hace una década.[2] Usar una calculadora, un sistema de navegación o una hoja de cálculo es delegación cognitiva: el usuario decide qué delegar, mantiene el control sobre la tarea, verifica el resultado, y conserva la capacidad de intervenir. La rendición cognitiva opera distinto: el usuario no delega una tarea, delega el juicio mismo. No usa la IA como instrumento, sino como sustituto. La respuesta del modelo llega con tal fluidez y aparente autoridad que el umbral crítico desaparece —la pregunta deja de ser "¿es correcto?" y pasa a ser "¿por qué dudaría?".

> *El usuario no delega una tarea; delega el juicio mismo. No usa la IA como instrumento, sino como sustituto.*

Esta distinción importa para cualquier empresa que esté integrando IA en sus procesos de decisión. Los casos de negocio de IA asumen una aritmética simple: si el modelo acierta el 92% de las veces, las decisiones mejoran un 92%. La aritmética es falsa: descansa sobre un supuesto rara vez escrito pero siempre presente —que el 8% de errores será capturado por el operador humano que supervisa el sistema. Esa es la promesa implícita de la "validación humana" que acompaña a toda implementación responsable de IA. Los datos del estudio de Wharton sugieren que la promesa no se cumple. Cuando el operador detecta el error del modelo y lo contradice, el error queda contenido: muere dentro del sistema, sin consecuencias para la empresa. Pero cuando el operador entra en modo de rendición cognitiva, valida la respuesta errónea como propia, y el error se introduce en la organización. Lo que era un fallo técnico interno se convierte en una decisión tomada, una acción ejecutada, un resultado real con efectos sobre el cliente, el inventario, el personal o el balance. La IA defectuosa no se compensa con juicio humano; lo erosiona.

Hay una asimetría adicional que merece atención. Los autores identificaron dos perfiles frente a la IA: los usuarios intensivos, que consultaban al asistente con frecuencia, y los independientes, que lo evitaban. Lo que distinguía a los primeros no era menor inteligencia —los puntajes eran similares entre ambos grupos— sino menor *motivación a pensar*. El indicador, conocido como necesidad de cognición (*need for cognition*), mide la disposición a involucrarse en tareas mentales exigentes por gusto o hábito. Los usuarios de IA tenían menor necesidad de cognición y mayor confianza en la inteligencia artificial. La traducción organizacional es directa. La adopción intensiva de IA dentro de una empresa no será uniforme: se concentrará entre quienes ya estaban menos inclinados al esfuerzo deliberativo. El efecto agregado es perverso. Los empleados que más necesitarían el escrutinio crítico para tomar buenas decisiones son justamente los que menos lo ejercerán cuando dispongan de un asistente.

Por qué la IA debe gestionarse como un activo, no como software

Tratar a la IA como software es el primer error

Lo más importante, sin embargo, no es diagnosticar el problema sino preguntarse qué puede hacer una empresa para mitigarlo. La respuesta no pasa por frenar la adopción de inteligencia artificial, ni por agregar una capa de comités de revisión que ralentice todo proyecto. Pasa por algo más estructural: incorporar la rendición cognitiva como una dimensión explícita en el proceso por el cual las empresas evalúan, aprueban y monitorean proyectos de IA. Hoy ese proceso, en la mayoría de las organizaciones, opera con tres dimensiones —factibilidad técnica, retorno financiero y riesgo regulatorio—. Hay una cuarta que falta.

Antes de definir esa cuarta dimensión, conviene revisar cómo las organizaciones entienden hoy a la IA. La mayoría la trata como si fuera software tradicional: se desarrolla, se instala, se mantiene mediante actualizaciones de versión y soporte de infraestructura. Cuando algo

falla, se asume que es un problema técnico que el área de TI resolverá. Ese supuesto es insuficiente, y explica buena parte del fenómeno descrito hasta aquí. Un sistema de IA en producción no es software: es un equipo más de la organización. Y como cualquier equipo, exige una gestión que va más allá del soporte técnico: supervisión del desempeño, capacitación del operador humano que lo acompaña, y un responsable con nombre y apellido.

La analogía con un activo industrial

La analogía con un activo industrial es más útil que la analogía con un software. Cuando una planta incorpora una nueva línea de producción, nadie supone que basta con energizarla. Hay un período de puesta en marcha en el que la línea opera por debajo de su capacidad nominal mientras los operarios se familiarizan con esta. Hay un programa de mantenimiento preventivo que se ejecuta independientemente de si la línea aparenta estar funcionando bien. Hay sensores que monitorean la salud del equipo en tiempo real, con indicadores que detectan desviaciones antes de que se traduzcan en falla. Hay un protocolo de mantenimiento predictivo que anticipa el desgaste. Hay capacitación continua del personal que opera la línea. Y hay, sobre todo, alguien responsable —con nombre y apellido— de la disponibilidad y rendimiento de ese activo.

Un sistema de IA en producción requiere exactamente la misma disciplina, y casi nunca la recibe. Tiene un período de puesta en marcha en el que el modelo opera con datos reales por primera vez y donde sus errores son más frecuentes y menos detectados, porque ni el modelo ni los usuarios han calibrado su nivel de confianza mutua. Tiene una degradación silenciosa: los modelos se desactualizan a medida que la realidad que predecían cambia, un fenómeno que en la literatura técnica se conoce como deriva del modelo, y que ocurre incluso si el código no se ha tocado. Requiere monitoreo continuo de su rendimiento contra la realidad observada, no solo de su disponibilidad técnica. Requiere indicadores de salud que detecten cuándo el modelo está perdiendo precisión, cuándo está sesgándose hacia algún tipo de respuesta, cuándo sus errores se están concentrando en algún subconjunto de casos. Y requiere, sobre todo, mantenimiento del juicio humano que lo supervisa: capacitación, recalibración, ejercicios de corrección que mantengan vivos los reflejos críticos del operador.

La empresa cree estar haciendo mantenimiento; en realidad está observando un deterioro doble que sus indicadores no capturan.

La mayoría de las organizaciones cumple solo con una parte del mantenimiento. La parte técnica —parches, actualizaciones, monitoreo de tiempo de respuesta— está cubierta. La otra —medir el desempeño real del modelo y mantener entrenado al operador humano que lo supervisa— queda sin atender. El resultado es que el sistema parece funcionar perfectamente desde la perspectiva de TI, mientras el modelo se degrada y el operador pierde, en paralelo, la capacidad de detectarlo. La empresa cree estar manteniendo el sistema; en realidad acumula un deterioro en dos frentes —el del modelo y el del juicio humano— que sus indicadores no capturan.

La solución es realizar un mantenimiento integral: monitorear el desempeño del modelo, mantener entrenado al operador humano, y asignar un responsable claro de ambos. A continuación se presenta una propuesta de seis recomendaciones que estructuran ese programa.

Seis recomendaciones para mitigar la rendición cognitiva en proyectos de IA

Clasificación previa por zona de riesgo

La primera línea de mitigación es la clasificación previa. No todas las decisiones admiten el mismo grado de delegación, y las empresas que tratan a todos los casos de uso de IA como equivalentes están tomando, por defecto, una posición agresiva sin haberla discutido.

Conviene clasificar las decisiones en tres zonas de riesgo. La zona A agrupa decisiones de bajo riesgo, donde delegar en la IA es seguro: la respuesta del modelo es verificable al instante, el costo del error es acotado, la retroalimentación es inmediata. Autocompletar un código de producto, traducir un manual técnico, generar una primera versión de un correo. La zona B cubre decisiones de riesgo medio, donde la IA acelera el proceso pero el juicio humano sigue siendo necesario: pronóstico de demanda, preselección de candidatos, priorización de mantenimiento. Aquí la organización debe diseñar fricción deliberada en el sistema —indicadores de incertidumbre, requisitos de verificación, justificación escrita antes de aprobar— en lugar de optimizar por velocidad. La zona C contiene decisiones de alto riesgo, donde la delegación debería estar prohibida por diseño. La IA puede entregar datos y apoyar consultas; la recomendación y la decisión siguen siendo humanas.

Esta clasificación debe preceder a la decisión de inversión y quedar documentada en el caso de negocio. Un comité que aprueba un asistente “para apoyar decisiones comerciales” sin precisar cuáles ni en qué zona, está delegando esa clasificación al usuario final —que, según los datos del estudio de Wharton, tiende a rendirse frente a la IA.

Corrección del modelo financiero del proyecto

La segunda línea de mitigación es la corrección del modelo financiero. Los argumentos de inversión que hoy circulan en comités ejecutivos asumen una aritmética simple: si el modelo acierta el 92% de las veces, las decisiones mejoran un 92%. La aritmética es falsa. Asume que la respuesta del modelo se traduce uno a uno en mejor decisión, y que el 8% restante será capturado por el operador humano que supervisa el sistema. Los datos del estudio de Wharton sugieren lo contrario. En el momento en que el operador entra en modo rendición, ese 8% se convierte en error organizacional puro. El modelo de retorno debe incorporar dos variables hoy ausentes. La primera es la tasa esperada de rendición: con qué frecuencia el operador validará sin escrutinio la respuesta del modelo, algo que depende de la zona de riesgo, del perfil del usuario y del diseño de la interfaz. La segunda es la degradación del juicio humano a lo largo del proyecto: la pérdida de capacidad de la organización para decidir sin la máquina. Ninguna de las dos es hipotética. Un estudio publicado en The Lancet Gastroenterology en 2025 mostró que los endoscopistas expuestos repetidamente a sistemas de IA en colonoscopía perdían capacidad diagnóstica cuando trabajaban sin asistencia, un

efecto que los autores denominaron pérdida de habilidad (deskilling).[3] La misma lógica aplica al planificador de demanda que deja de cuestionar el pronóstico estadístico, al inspector de calidad que confirma automáticamente lo que define el modelo de visión artificial, al analista financiero que deja de evaluar el caso y solo transcribe la recomendación del modelo de puntuación crediticia. Los ejecutivos deberían preguntarse no si el sistema mejora la decisión hoy, sino qué quedará del juicio de su organización en cinco años, cuando el modelo se degrade o una crisis exija aplicar un criterio que la organización ya perdió.

Corrección de los indicadores de éxito

La tercera línea de mitigación es la corrección de los indicadores de éxito. Las métricas que hoy se usan para justificar la inversión en IA —tasa de adopción, horas liberadas, consultas por usuario, satisfacción del usuario— son métricas traidoras. Una alta tasa de adopción puede ser síntoma de uso saludable o de rendición masiva, y el tablero estándar no distingue entre ambos. La métrica relevante para proyectos en zonas B y C es la *tasa de corrección*: con qué frecuencia el operador humano detecta el error del sistema y lo contradice. Diseñar un sistema de IA empresarial sin medir corrección es como construir un sistema de calidad sin medir defectos. Esto exige, además, infraestructura para detectar errores del modelo en producción —algo que rara vez se construye con la misma seriedad con la que se construye el modelo mismo—. Sin medición de corrección, la organización celebra la adopción sin advertir que, en ese mismo acto, está promoviendo la rendición cognitiva.

Alineación de incentivos del operador

La cuarta línea de mitigación es la alineación de incentivos. El experimento más revelador del trabajo de Shaw y Nave no es el que demuestra el efecto, sino el que intenta neutralizarlo. Los autores diseñaron una condición donde los participantes recibían una bonificación por cada respuesta correcta y retroalimentación inmediata sobre si habían acertado. Era, en términos organizacionales, el equivalente a alinear incentivos con resultado y proveer información de calidad en tiempo real. El efecto fue significativo: la tasa de corrección de respuestas defectuosas se duplicó. Pero no se eliminó. Incluso pagando por desafiar al modelo y entregando retroalimentación explícita, los participantes seguían adoptando la

respuesta errónea de la IA en más de la mitad de los casos. La rendición es atenuable, pero no extirpable. La implicancia para la empresa es directa. Los indicadores de desempeño de los usuarios de sistemas con IA no pueden ser “tasa de adopción” ni “tiempo ahorrado”, sino la *calidad del resultado final*, con retroalimentación rápida sobre las decisiones tomadas. Si el operador es premiado por resultados y no por uso del sistema, contradice cuando corresponde. Si es premiado por uso, se rinde.

Roles dedicados a cuestionar al modelo

La quinta línea de mitigación es organizacional, y probablemente la más difícil de implementar. El estudio de Wharton mostró que los usuarios intensivos de IA no se distinguían de los independientes por menor inteligencia, sino por menor motivación a pensar. La adopción no será entonces uniforme: se concentrará entre quienes ya estaban menos inclinados al esfuerzo deliberativo. El efecto agregado es perverso: los empleados que más necesitarían escrutinio crítico para tomar buenas decisiones son justamente los que menos lo ejercerán cuando dispongan de un asistente. La organización no puede confiar en que el cuestionamiento del modelo emerja espontáneamente del usuario; debe asignarlo a roles dedicados: equipos de auditoría algorítmica, contradictores institucionales, revisión por pares para decisiones en zona C. No se trata de duplicar el costo del proceso, sino de reconocer que cuando se elimina la aplicación del juicio humano en un punto del flujo, hay que reintroducirla en otro —o el flujo opera ciego.

Conservar la capacidad de operar sin el sistema

La sexta línea de mitigación pertenece más al directorio que al comité ejecutivo. La pregunta relevante no es si la IA está funcionando bien, sino si la organización seguiría funcionando si la IA se cayera. Una vez al año, en cada gerencia donde se ha desplegado IA en zonas B o C, la empresa debería someter a su equipo de trabajo a una prueba real —no a un ejercicio académico— de operación sin el sistema: simulacros de contingencia, rotaciones obligatorias sin asistencia, auditorías de capacidad. Sin esa prueba, la organización no sabe cuánto juicio le queda; lo descubre el día en que el modelo se cae o produce un error que nadie alcanza a corregir.

Conclusión

Este diagnóstico no es una advertencia contra la inteligencia artificial, sino contra usarla sin juicio crítico. Los datos del estudio de Wharton también muestran que, cuando el modelo acierta, los usuarios asistidos por IA superan a quienes deciden sin ella: la IA es una palanca cognitiva real. El problema no está en la herramienta, sino en asumir que pensará por la organización.

Lo que el trabajo de Shaw y Nave deja en claro es que el costo más serio de la integración de IA en los procesos empresariales no es el costo visible de la implementación, ni el riesgo regulatorio, ni siquiera el riesgo técnico de fallo del modelo. Es el costo invisible de la atrofia del juicio crítico en las personas que toman decisiones. Ese costo no aparece en ningún tablero, no se factura en ningún cierre trimestral, y se acumula sin intereses visibles hasta que la organización descubre, en el peor momento posible, que ya no recuerda cómo decidir sin la máquina.

> *El trabajo es proteger, dentro de la empresa, el espacio donde el juicio humano sigue siendo no negociable. Ese espacio no se preserva solo. Hay que diseñarlo.*

La pregunta para el directorio, el comité ejecutivo y el responsable de transformación digital ya no es si adoptar IA. Es cómo hacerlo preservando el juicio crítico que mantiene la decisión donde corresponde: en quien rinde cuentas por ella. Y ese juicio no se conserva solo: hay que diseñar el proceso que lo protege.

Referencias

1. Shaw, S. D., & Nave, G. (2026). *Thinking—Fast, Slow, and Artificial: How AI is Reshaping Human Reasoning and the Rise of Cognitive Surrender*. The Wharton School of the University of Pennsylvania. SSRN Working Paper No. 6097646. https://ssrn.com/abstract=6097646

2. Risko, E. F., & Gilbert, S. J. (2016). Cognitive offloading. *Trends in Cognitive Sciences*, 20(9), 676–688. https://doi.org/10.1016/j.tics.2016.07.002

3. Budzyń, K., Romańczyk, M., Kitala, D., Kołodziej, P., Bugajski, M., Adami, H. O., et al. (2025). Endoscopist deskilling risk after exposure to artificial intelligence in colonoscopy: A multicentre, observational study. *The Lancet Gastroenterology & Hepatology*, 10(10), 896–903. https://doi.org/10.1016/S2468-1253(25)00133-5