Capacidades para la era de la IA agéntica: gobernanza epistémica y co-inteligencia humano-máquina

Capabilities for the age of agentic AI: epistemic governance and human–machine Co-Intelligence

Áurea Rodríguez López
Especialista en estrategias de innovación, escritora, estudiante de doctorado, UIC, directora corporativa Eurecat

Resumen:
La transición desde la automatización clásica hacia sistemas de inteligencia híbrida se acelera con la irrupción de la IA agéntica: arquitecturas capaces de planificar, decidir y ejecutar flujos de trabajo mediante el uso de herramientas y datos en bucle. Este cambio desplaza el foco desde la pregunta por la sustitución o la mera "ayuda" al trabajo hacia un problema más fino: qué capacidades deben adquirir personas, equipos y organizaciones para delegar, supervisar y aprender con agentes cada vez más competentes sin perder criterio, trazabilidad ni control. El artículo propone un marco de "capacidades para la co-inteligencia" que integra: (i) capacidades cognitivas (formulación de pro-blemas, vigilancia epistémica, verificación y calibración), (ii) capacidades socio-técnicas (diseño de instrucciones, orquestación de herra-mientas, evaluación continua y recuperación ante fallos) y (iii) capacidades de gobernanza (gestión de datos, registro de decisiones, auditoría y rendición de cuentas). A partir de evidencia empírica sobre productividad y nivelación de habilidades, y de la literatura sobre offloading cognitivo, se argumenta que la adopción madura de la IA agéntica depende menos de "modelos más inteligentes" que de prácticas repetibles que conviertan la delegación en aprendizaje y reduzcan la exposición a errores encadenados, inyección de instrucciones y sobre-autonomía. El texto concluye con recomendaciones para el diseño organizativo y la formación: alfabetización agéntica, evaluación como hábito, arquitectura de permisos, y cultura de trazabilidad. La regulación, en particular, las exigencias de supervisión humana del AI Act, se interpreta como un marco habilitador que obliga a operacionalizar capacidades, no como un sustituto de ellas.

Palabras clave:
IA agéntica; Inteligencia híbrida; Capacidades; Gobernanza de IA; Co-inteligencia humano-máquina.

Abstract:
The shift from traditional automation to hybrid intelligence is accelerating with the rise of agentic AI: architectures that can plan, decide, and execute workflows by looping over tools and data. This transformation reframes the debate from substitution versus augmentation to a more precise question: which capabilities do individuals, teams, and organizations need in order to delegate, supervise, and learn with increasingly competent agents without losing judgment, traceability, or control? This article proposes a "co-intelligence capability" framework integrating (i) cognitive capabilities (problem framing, epistemic vigilance, verification and calibration), (ii) socio-technical capabilities (instruction design, tool orchestration, continuous evaluation, and failure recovery), and (iii) governance capabilities (data management, decision logging, auditing, and accountability). Drawing on empirical evidence on productivity and skill leveling, and on the cognitive offloading literature, it argues that mature adoption of agentic AI depends less on "smarter models" than on repeatable practices that turn delegation into learning while reducing exposure to cascading errors, instruction injection, and excessive autonomy. The article concludes with recommendations for organizational design and training: agentic literacy, evaluation as a routine, permission architecture, and a culture of traceability. Regulation, particularly the human oversight requirements in the EU AI Act, is interpreted as an enabling scaffold that forces capability operationalization rather than replacing it.

Keywords:
Agentic AI; Hybrid intelligence; Capabilities; AI governance; Human–machine co-intelligence.

1. Introducción

En apenas dos años, la conversación pública sobre la inteligencia artificial (IA) ha pasado de la "automatización" a la "aumentación" y ahora, a la "colaboración". La IA generativa, inicialmente utilizada como asistente conversacional o herramienta de redacción, está dando paso a sistemas más autónomos que planifican y ejecutan acciones: consultan fuentes, llaman a herramientas, actualizan registros, crean artefactos y devuelven resultados con resultados superiores a los humanos en algunos casos. Este giro se suele describir como IA agéntica, un término que alude a la capacidad de un sistema para orientar su conducta hacia un objetivo a lo largo de varios pasos, interactuando con un entorno mediante herramientas y feedback.

Este cambio se entiende mejor si se adopta el paradigma de la inteligencia híbrida: configuraciones socio-técnicas en las que agentes humanos y algorítmicos co-producen resultados y donde el rendimiento emerge de su interacción, no de cada parte por separado[1]. En lugar de preguntar "qué tareas hace la IA", conviene preguntar "qué tareas se vuelven delegables" y "qué tareas se vuelven críticas" cuando la IA asume partes del flujo. La IA agéntica empuja esta lógica porque deja de limitarse a sugerir contenido, sino que puede explorar el espacio de soluciones, seleccionar herramientas y ejecutar pasos intermedios, de modo que la interacción humana se desplaza hacia el diseño del encargo, la autorización de acciones y la verificación.

Para evitar confusiones conceptuales, hablaremos de co-inteligencia humano-máquina para referirnos a procesos en los que la IA participa en la generación de hipótesis, en la búsqueda de evidencias y en la redacción o cálculo, mientras que el humano conserva responsabilidad sobre los objetivos, la selección de evidencias decisivas y la decisión final. Esta definición se alinea con la idea de la IA como "copiloto" y con recomendaciones industriales que insisten en que el valor depende del diseño de la interacción, de la calidad de los datos y de la capacidad de evaluación, no solo del tamaño del modelo.

[1] Dominik Dellermann, Philipp Ebel, Matthias Söllner y Jan Marco Leimeister, "Hybrid Intelligence," *Business & Information Systems Engineering* 61, no. 5 (2019): 637–643, https://doi.org/10.1007/s12599-019-00595-2

Esta evolución no convierte automáticamente a la máquina en un sujeto moral o jurídico; "agencia" aquí es un rasgo funcional, no una atribución de personalidad. Sin embargo, sí altera la economía cognitiva del trabajo del conocimiento. Cuando el sistema deja de ser una calculadora avanzada y empieza a ejecutar tareas compuestas, con acceso a herramientas y a datos, el papel humano se desplaza: menos teclear y más definir objetivos, fijar restricciones, verificar resultados, decidir excepciones y aprender del desempeño. La diferencia no es solo de grado (más automatización) sino de tipo: cambia la unidad de trabajo delegable, que ya no es una tarea discreta, sino un proceso.

En este contexto, el debate "sustitución versus aumento" se queda corto por dos motivos. Primero, porque la sustitución no depende solo de la capacidad del modelo, sino del ensamblaje socio-técnico (datos, herramientas, procesos, incentivos) que hace que una tarea sea delegable de forma fiable. Segundo, porque el aumento puede degradar capacidades humanas si se convierte en externalización sistemática del pensamiento (cognitive offloading) sin prácticas de verificación y aprendizaje. El problema ya no es si la IA "ayuda", sino si ayuda de un modo que preserve o incluso fortalezca el criterio.

La hipótesis central de este artículo es que la adopción eficaz de IA agéntica requiere un conjunto de capacidades específicas, distribuibles entre individuos, equipos y organización, que hacen posible la co-inteligencia: la construcción conjunta de conocimiento y decisiones por agentes humanos y algorítmicos. La propuesta implica un desplazamiento analítico: desde derechos y obligaciones como punto de partida hacia capacidades como unidad de diseño. El derecho sigue siendo relevante, pero en un sentido instrumental: establece un suelo de garantías y un lenguaje de responsabilidad, pero no resuelve por sí mismo qué prácticas, competencias y artefactos deben existir para que la supervisión sea efectiva. En un sistema agéntico, "tener un humano en el bucle" no es una propiedad; es el resultado de un diseño.

2. De la IA generativa a la IA agéntica: qué cambia

Para entender el giro agéntico conviene distinguir entre tres niveles de integración. En el nivel 1, la IA actúa como motor de producción: genera texto, código o resúmenes a partir de una instrucción única. En el nivel 2, la IA se integra en un flujo guiado: se encadena con recuperación de información, plantillas, validaciones y reglas; el recorrido está determinado por el software. En el nivel 3, el sistema decide el recorrido: el modelo gestiona la ejecución del flujo de trabajo, elige herramientas y controla la iteración (plan-acción-observación-replanificación) hasta cumplir un criterio de parada. Este tercer nivel se aproxima a lo que la literatura industrial llama "agentes" o "sistemas agénticos"[2].

Dos rasgos distinguen este nivel 3. El primero es la autonomía operacional: el agente decide qué pasos dar, no solo cómo redactar un paso ya definido. El segundo es la ampliación del "espacio de acción" mediante herramientas: el agente puede consultar una base de datos, ejecutar código, escribir en un CRM o lanzar una búsqueda. La combinación de autonomía y herramientas crea un sistema con capacidad de actuación, por lo que el error ya no es únicamente semántico (una respuesta incorrecta) sino operativo (una acción incorrecta).

2.1. Arquitectura agéntica: planificar, actuar, observar y corregir

Los agentes operan, típicamente, en bucles. Una formulación simple es plan-act-observe: el sistema propone un plan, ejecuta una acción (por ejemplo, llamar a una herramienta), observa el resultado y decide el siguiente paso. En versiones más robustas, el bucle incorpora una fase de crítica o reflexión: el agente evalúa si el resultado satisface el criterio y, si no, re-intenta o replantea el plan. Este patrón, documentado en guías de implementación, explica por qué los agentes pueden resolver tareas abiertas sin que el programador especifique todos los pasos: la búsqueda de una trayectoria adecuada se hace en tiempo de ejecución.

Este diseño tiene implicaciones directas para la supervisión. En un sistema de una sola respuesta, la supervisión se concentra en el output. En un sistema agéntico, la supervisión puede situarse en el plan (validar antes de actuar), en las herramientas (limitar permisos), en los checkpoints (revisiones parciales) o en el criterio de parada (cuándo el agente debe considerar la tarea "terminada"). La pregunta "¿dónde pongo al humano?" se vuelve una decisión de ingeniería: depende del coste de verificación y del daño potencial de una acción incorrecta.

[2] Véanse, entre otras, OpenAI, "A Practical Guide to Building Agents" (documento PDF), consultado el 31 de diciembre de 2025, https://cdn.openai.com/business-guides-and-resources/a-practical-guide-to-building-agents.pdf; y Anthropic, "Building Effective Agents," 19 de diciembre de 2024, https://www.anthropic.com/research/building-effective-agents.

2.2. Memoria, contexto y límites: por qué "más inteligencia" no elimina el riesgo

Los modelos actuales trabajan con ventanas de contexto finitas y con memorias externas diseñadas por el sistema (por ejemplo, un registro de trabajo o una base documental). En IA agéntica, la memoria es una pieza crítica: determina qué información se conserva, cómo se resume y qué se olvida. Las decisiones de memoria afectan tanto a la calidad como a la seguridad: un agente que "recuerda" demasiado puede exponer datos; uno que "olvida" demasiado puede repetir errores o perder restricciones. Además, la capacidad de un agente para operar en entornos reales depende de herramientas que no son perfectas: buscadores con sesgos, bases de datos incompletas, APIs con errores, documentos ambiguos. El punto de vista de capacidades ayuda a evitar el espejismo de la "superinteligencia práctica": lo que importa es el sistema completo, incluido el diseño de herramientas, permisos, registros y métricas.

Desde una perspectiva de diseño, el salto al nivel 3 reorganiza la arquitectura del trabajo en torno a tres componentes: modelo, herramientas e instrucciones.[3] El modelo aporta capacidad de razonamiento y planificación; las herramientas permiten actuar; y las instrucciones definen objetivos, límites, criterios de calidad y condiciones de parada. El desempeño no depende solo del modelo "más inteligente", sino del acoplamiento entre estos componentes: herramientas mal diseñadas o mal documentadas aumentan errores; instrucciones ambiguas amplifican desviaciones; y un modelo muy capaz sin permisos adecuados puede ser más riesgoso que uno menos capaz con guardarraíles.

La IA agéntica introduce además una nueva unidad de evaluación. En IA generativa, se evalúa una respuesta (un texto) respecto a un criterio. En IA agéntica, se evalúa una trayectoria: una secuencia de decisiones y acciones. Esto obliga a pensar en métricas de proceso: número de herramientas llamadas, tasa de errores, tiempo por paso, cobertura de evidencias, y calidad del registro. En la práctica, este enfoque desplaza el control desde el "resultado final" hacia la trazabilidad de la ejecución.

Un riesgo específico de los sistemas agénticos es la "agencia excesiva". Si el agente puede ejecutar acciones con pocos frenos, un error de interpretación o una instrucción maliciosa (por ejemplo, *prompt injection* indirecta en un documento que el agente lee) puede traducirse en acciones no deseadas. Marcos de seguridad recientes para aplicaciones con modelos de lenguaje identifican la agencia excesiva y la inyección de instrucciones como riesgos centrales.[4] Por eso, la seguridad deja de ser un asunto exclusivamente técnico y se convierte en una práctica operativa: delimitar permisos, aislar entornos, diseñar herramientas ergonómicas para agentes y crear registros auditables. Un ejemplo ayuda. Supongamos un agente encargado de preparar un informe de diligencia debida sobre un proveedor. En un enfoque no agéntico, el profesional pide un resumen, revisa y corrige. En un enfoque agéntico, el agente: (i) consulta bases internas, (ii) busca información pública, (iii) extrae indicadores, (iv) redacta un borrador, (v) genera una matriz de riesgos, y (vi) propone recomendaciones. En cada paso puede equivocarse: confundir identidades, interpretar erróneamente una norma, o incorporar una instrucción maliciosa encontrada en una web. El resultado final puede parecer impecable; el problema puede estar en la trayectoria. La pregunta clave es: ¿dispone el usuario de medios para inspeccionar esa trayectoria sin rehacer todo el trabajo?

Esa pregunta conduce a la tesis del artículo: operar con IA agéntica exige capacidades específicas que hacen viable la inspección, la verificación y el aprendizaje. Sin ellas, la organización puede obtener velocidad a corto plazo a costa de fragilidad a largo plazo.

3. Impactos en el trabajo del conocimiento: productividad, nivelación y *offloading*

La evidencia empírica sobre IA generativa en entornos de trabajo sugiere incrementos de productividad y cambios en la distribución de habilidades. Un estudio de referencia en un entorno de atención al cliente reporta mejoras medias relevantes y heterogéneas, con mayores beneficios para perfiles menos experimentados, y ganancias más modestas, o incluso pequeñas caídas de calidad, para perfiles expertos.[5] Este patrón se interpreta como "nivelación": el sistema difunde

[3] OpenAI, "New Tools for Building Agents," 11 de marzo de 2025, https://openai.com/index/new-tools-for-building-agents/.

[4] OWASP Foundation, "OWASP Top 10 for Large Language Model Applications (Version 1.1)," consultado el 31 de diciembre de 2025, https://owasp.org/www-project-top-10-for-large-language-model-applications/.

[5] Erik Brynjolfsson, Danielle Li y Lindsey Raymond, "Generative AI at Work," *The Quarterly Journal of Economics* 140, n.º 2 (2025): 889–942, https://doi.org/10.1093/qje/qjae044.

buenas prácticas, reduce la brecha entre novatos y expertos y acelera el aprendizaje inicial.

Esta nivelación no es un detalle: reconfigura la idea de "competencia" en el trabajo del conocimiento. Si un agente puede redactar, sintetizar y argumentar con un nivel razonable, el valor diferencial humano se desplaza hacia tareas menos visibles: definir el problema, seleccionar evidencias, negociae, detectar supuestos erróneos, y decidir bajo incertidumbre. En suma, se desplaza desde la producción hacia el juicio.

Este desplazamiento tiene dos efectos prácticos. Primero, hace que la experiencia sea menos visible: el experto aporta valor no tanto por producir un texto impecable, sino por saber qué preguntar, qué no creer y qué evidencia exigir. Segundo, cambia el "punto de fallo" de los procesos. Si antes el fallo típico era una mala redacción o un cálculo erróneo, ahora el fallo típico es un objetivo mal especificado o un supuesto no detectado que se amplifica a lo largo del flujo agéntico.

Desde el punto de vista organizativo, esto implica que la adopción de IA no es solo una cuestión de licencias o de herramientas, sino de rediseño de prácticas de calidad y de las propias organizaciones. Las organizaciones con mayor madurez tienden a formalizar criterios: definen qué decisiones exigen evidencia primaria, qué tareas requieren revisión por pares, qué umbrales activan escalado y qué métricas se monitorean (tasa de correcciones, incidencias de datos, retrabajo). La IA agéntica acelera la necesidad de estas prácticas.

Además, la nivelación de habilidades puede coexistir con una "polarización" de capacidades. Mientras las tareas de producción se democratizan, aumentan los retornos de las capacidades de verificación, orquestación y gobernanza. Esto sugiere una reconfiguración de la formación: ya no basta con enseñar a "hacer" (redactar, resumir), sino a "asegurar" (verificar, auditar) y a "diseñar" (definir objetivos, restricciones y evaluaciones).

Sin embargo, la misma lógica que facilita el aprendizaje puede favorecer la dependencia. La literatura sobre la descarga (*offloading*) cognitivo describe cómo las personas delegan memoria, cálculo y razonamiento a artefactos externos (desde una libreta hasta internet) y cómo esa delegación puede ser adaptativa o perjudicial según el contexto.[6] Con IA generativa, la externalización es más profunda: el sistema no solo almacena o busca, sino que produce inferencias, sugiere decisiones y completa razonamientos. Además, la interfaz conversacional hace que la delegación parezca diálogo: el usuario siente que "piensa con" el sistema incluso cuando está aceptando resultados sin reconstruirlos.

Tres mecanismos explican por qué la descarga cognitiva puede erosionar capacidades. Primero, reduce la práctica deliberada: si el sistema escribe o calcula por nosotros, practicamos menos la habilidad subyacente. Segundo, altera la memoria: cuando sabemos que un recurso externo está disponible, codificamos menos información interna (el "efecto Google"). Tercero, cambia la metacognición: la facilidad de obtener respuestas reduce la calibración del propio conocimiento y puede aumentar la confianza injustificada. Estos efectos no son inevitables, pero emergen cuando la organización incentiva la velocidad sin crear contrapesos.

La IA agéntica intensifica el dilema. Cuando el agente encadena pasos y devuelve un resultado final, el usuario ve menos del proceso y, por tanto, tiene menos oportunidades de detectar errores. La "fluidez" del resultado induce una ilusión de corrección: suena competente, luego debe estarlo. Este fenómeno se agrava en tareas con verificación costosa o ambigua (estrategia, análisis normativo, diagnóstico organizativo), donde no existen pruebas automáticas equivalentes a un test unitario.

En contraste, hay dominios donde el enfoque agéntico puede fortalecer capacidades: programación con tests, análisis con datos reproducibles, o investigación con trazabilidad de fuentes. La diferencia es la verificabilidad. Cuando se puede verificar sin rehacer el trabajo, se puede confiar sin fe. Por eso, la co-inteligencia madura se parece a la ingeniería: produce resultados y, al mismo tiempo, genera evidencia de que el resultado es bueno.

La pregunta relevante, por tanto, no es si la IA aumenta la productividad, en muchos casos lo hace, sino bajo qué condiciones esa productividad se convierte en aprendizaje y no en desentrenamiento. Ahí es donde el enfoque de capacidades aporta ventaja: permite describir prácticas concretas para obtener valor sin perder criterio. En lugar de pedir a las personas que "confíen menos", se trata de diseñar rutinas, herramientas y métricas que hagan la verificación viable y, por tanto, incentivada.

[6] Evan F. Risko y Sam J. Gilbert, "Cognitive Offloading," *Trends in Cognitive Sciences* 20, n.º 9 (2016): 676–688; Betsy Sparrow, Jenny Liu y Daniel M. Wegner, "Google Effects on Memory: Cognitive Consequences of Having Information at Our Fingertips," *Science* 333, n.º 6043 (2011): 776–778.

4. Un marco de capacidades para la co-inteligencia

Proponemos siete familias de capacidades que, en conjunto, permiten operar con IA agéntica de forma fiable. No son rasgos individuales "de talento": son combinaciones de conocimientos, prácticas y artefactos que pueden distribuirse entre personas, equipos, procesos y tecnología. En sistemas híbridos, las capacidades se co-construyen: lo que hace posible el juicio humano no es solo el "cerebro", sino el acceso a evidencias, registros, contra-ejemplos y mecanismos de control. La figura conceptual es una "superficie de control": cuanto más autonomía se delega, más importante es disponer de superficies donde intervenir.

Antes de detallar las familias de capacidades, conviene formular cuatro principios de diseño que las atraviesan:

(1) Verificabilidad antes que elocuencia. Un agente útil no es el que "suena" convincente, sino el que produce evidencias y se deja comprobar. En tareas donde no existe verificación razonable, la autonomía debe ser baja.

(2) Trazabilidad como condición del aprendizaje. La organización solo mejora el sistema si puede reconstruir qué ocurrió. Por eso, el registro (prompts, herramientas, datos, decisiones) no es burocracia: es memoria institucional.

(3) Autonomía graduada. No hay un único modo de uso. La autonomía debe ajustarse al riesgo y a la reversibilidad de la acción. Un agente puede tener autonomía alta en borradores y baja en aprobaciones o comunicaciones externas.

(4) Separación de roles. La co-inteligencia se fortalece cuando se separa generación y crítica. Un agente genera hipótesis; otro evalúa. Un humano decide. Esta separación reduce sesgos y favorece la calibración.

Estos principios pueden traducirse en una lista operativa de preguntas, útil como checklist al desplegar un agente:

- ¿Qué decisión o resultado produce el agente y qué daños causaría un error
- ¿Qué evidencias debe aportar para que un humano pueda verificarlo sin rehacer todo el trabajo
- ¿Qué acciones puede ejecutar y qué permisos requiere? ¿Hay acciones irreversibles?
- ¿Qué registros se guardan para permitir auditoría y aprendizaje?

Con estos principios en mente, pasamos a las capacidades.

4.1. Capacidad de formulación de problemas y objetivos

En contextos agénticos, lo difícil no es "pedir" una respuesta, sino especificar qué cuenta como éxito. Esta capacidad incluye: delimitar el problema, identificar restricciones (legales, éticas, presupuestarias), descomponer objetivos en subtareas y traducir criterios de calidad en instrucciones verificables. Se parece a la ingeniería de requisitos, pero aplicada a tareas cognitivas y socio-organizativas.

Una práctica básica es la especificación de salida: antes de ejecutar, describir el formato, el nivel de evidencia y los supuestos permitidos. Por ejemplo: "Redacta un informe de 1.500 palabras con tres opciones estratégicas. Para cada opción, incluye (a) supuestos explícitos, (b) evidencia citada, (c) riesgos y mitigaciones, y (d) condiciones de reversibilidad". Esta especificación reduce alucinaciones porque obliga a anclar el texto en estructura y evidencia.

Otra práctica es el "contrato de incertidumbre": pedir al agente que señale qué partes dependen de supuestos no verificados y qué datos faltan. En IA agéntica, esto se extiende a la planificación: el agente debe proponer un plan y pedir confirmación cuando el plan implique acciones irreversibles o acceso a datos sensibles. La formulación, así, no es solo una instrucción; es un diálogo de diseño.

4.2. Capacidad de delegación y orquestación

Delegar a un agente no es "ceder" la tarea: es asignar una sub-tarea con límites, permisos y checkpoints. Esta capacidad comprende: seleccionar herramientas adecuadas, definir ám-

bitos de acceso (principio de mínimo privilegio), usar patrones de orquestación (por ejemplo, "orquestador-trabajadores" o "evaluador-optimizador") y coordinar múltiples agentes especializados.[7]

En términos prácticos, implica construir un mapa de trabajo donde cada agente tiene: rol, objetivo, conjunto de herramientas autorizadas, criterio de parada y mecanismo de escalado. Un patrón útil es separar "agentes de generación" y "agentes de crítica". En escritura, por ejemplo, un agente produce un borrador y otro lo revisa con criterios explícitos. En investigación, un agente propone fuentes y otro valida que las fuentes son primarias y relevantes.

La delegación también exige comprender costes y latencias. Los agentes agénticos pueden ser caros: cada iteración consume recursos. Por eso, parte de la capacidad es saber cuándo no usar agentes. Un principio defendido en guías industriales es comenzar con soluciones simples y aumentar complejidad solo cuando demuestre mejorar resultados.[8] Delegar bien es también decidir qué delegar.

4.3. Capacidad de vigilancia epistémica y verificación

La vigilancia epistémica es la habilidad de evaluar la fiabilidad de una información o un razonamiento en función de su fuente, coherencia y plausibilidad. En humanos, opera a través de heurísticas; en co-inteligencia, debe complementarse con prácticas formales: trazabilidad de fuentes, contrastación con evidencia primaria y verificación de afirmaciones críticas. La literatura sobre comunicación y cognición destaca que la confianza no es global: es una evaluación situada y revisable[9].

En IA agéntica, esta capacidad se traduce en protocolos. Proponemos una regla operativa: toda afirmación "decisiva" (aquella que, si es falsa, cambia la decisión) debe estar respaldada por una fuente primaria o por un dato reproducible. Esto obliga a diseñar al agente para que cite fuentes y a diseñar herramientas para devolver contexto suficiente. Una cita sin contexto es una falsa trazabilidad.

La verificación puede ser humana o automática. En software, tests. En análisis de datos, reproducibilidad (scripts y datasets). En regulación, consulta de textos oficiales. En tareas blandas, triangulación (múltiples fuentes y perspectivas). El objetivo no es eliminar incertidumbre, sino reducirla y localizarla.

4.4. Capacidad de calibración y metacognición

Incluso con verificación, muchas tareas tienen incertidumbre residual. Calibrar consiste en ajustar la confianza subjetiva al desempeño real: saber cuándo el sistema suele fallar, en qué dominios es fiable y qué señales anticipan errores. Esta capacidad requiere métricas y feedback: sin medición, la confianza se convierte en opinión.

En agentes, la calibración se apoya en evals y en análisis de fallos similares a los de ingeniería de software. Guías recientes insisten en construir evaluaciones para herramientas y agentes y en analizar transcripciones para detectar patrones de confusión[10]. La organización que adopta IA agéntica debería mantener un repositorio de "casos de fallo", etiquetados por causa (datos faltantes, ambigüedad, herramienta mal definida, *prompt injection*, etc.) y por severidad.

La metacognición incluye también saber "qué no sé". Paradójicamente, un sistema muy elocuente puede ocultar ignorancia. Por eso, una práctica útil es pedir estimaciones probabilísticas o rangos, y comparar esas estimaciones con resultados reales. Con el tiempo, se construye una "tabla de calibración" por dominio: por ejemplo, en redacción de comunicaciones, el agente es fiable; en interpretación jurisprudencial reciente, requiere verificación intensa; en síntesis de políticas internas, depende de si tiene acceso al repositorio correcto.

[7] Anthropic, "Building Effective Agents.".

[8] OpenAI, "A Practical Guide to Building Agents."

[9] Dan Sperber et al., "Epistemic Vigilance," *Mind & Language* 25, n.º 4 (2010): 359–393.

[10] National Institute of Standards and Technology (NIST), *Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile* (NIST AI 600-1) (Gaithersburg, MD: NIST, 2024), https://doi.org/10.6028/NIST.AI.600-1.

4.5. Capacidad de resiliencia cognitiva frente a la descarga cognitiva

La descarga cognitiva no es mala en sí, lo es cuando sustituye la práctica necesaria para conservar pericia. Resiliencia cognitiva significa diseñar un uso de IA que preserve el entrenamiento: alternar modos (con y sin IA), practicar verificación activa (reconstruir el argumento con palabras propias), y mantener espacios de trabajo donde la persona produce sin ayuda para consolidar memoria y comprensión.

Proponemos tres técnicas simples. (i) Doble entrega: primero un resultado asistido y luego una versión "explicada" por la persona, donde se justifiquen supuestos y se expongan evidencias. (ii) Inversión de roles: pedir al agente que critique un razonamiento humano, pero no que lo sustituya; esto fomenta aprendizaje deliberado. (iii) "Ayuda tardía": en tareas de aprendizaje, retrasar el uso de IA hasta haber intentado resolver por cuenta propia. Estas prácticas mantienen la agencia cognitiva y reducen la tentación de aceptar sin comprender.

4.6. Capacidad de seguridad operativa en sistemas agénticos

Los agentes amplían la superficie de ataque porque consumen entradas externas y ejecutan acciones. La seguridad se convierte en competencia transversal: reconocer prompt injection, separar instrucciones del contenido, aislar herramientas, registrar acciones y limitar permisos. Marcos como el OWASP Top 10 para aplicaciones con modelos de lenguaje enumeran riesgos específicos (inyección de prompts, manejo inseguro de salidas, agencia excesiva) y ofrecen un vocabulario útil para traducir seguridad a checklist[11].

En entornos agénticos, el principio clave es la separación de dominios: el agente puede leer contenido no confiable, pero no debe tratarlo como instrucciones. Esto implica filtros, delimitadores, y políticas de "no ejecutar" sin confirmación humana para acciones irreversibles. También implica sandboxes: probar agentes en entornos aislados con datos sintéticos antes de permitir acceso a sistemas reales.

La seguridad incluye además el diseño de herramientas. Herramientas "ergonómicas" para agentes devuelven información de alto valor, no volúmenes masivos; usan nombres y parámetros claros; y evitan exponer identificadores opacos. La literatura de ingeniería agéntica destaca que herramientas bien diseñadas reducen alucinaciones y errores, porque el agente dispone de affordances más naturales[12]. En este sentido, la seguridad no es un parche: es un diseño de interfaz para sistemas no deterministas.

4.7. Capacidad de gobernanza: datos, trazabilidad y rendición de cuentas

La última familia de capacidades es organizativa. Incluye: gobierno de datos (calidad, licitud, minimización), trazabilidad de decisiones (logs, versionado de prompts, modelos y herramientas), auditoría y mecanismos de rendición de cuentas. En Europa, el AI Act incorpora obligaciones de supervisión humana para sistemas de alto riesgo que, en la práctica, exigen operacionalizar esta capacidad: no basta con "un humano en el bucle" si no tiene información, autoridad y tiempo para intervenir[13]. La gobernanza se concreta en artefactos. Por ejemplo: (i) registro de versiones de modelos y prompts, (ii) bitácora de acciones de herramientas, (iii) políticas de datos (qué se puede usar, cómo se anonimiza, cuánto se retiene), (iv) criterios para escalado a revisión humana, y (v) procedimientos de respuesta a incidentes. Estos artefactos permiten que el sistema sea auditable y, por tanto, mejorable. Sin ellos, los fallos se repiten porque no se aprende de forma institucional.

5. Gobernanza como infraestructura de capacidades: del "control" a la trazabilidad

En la adopción temprana de IA generativa, muchas organizaciones han confundido control con prohibición o con controles superficiales (por ejemplo, exigir que alguien "revise" al final). En IA

[11] OWASP Foundation, "OWASP Top 10 for Large Language Model Applications (Version 1.1)."

[12] Anthropic, "Engineering at Anthropic: Writing Tools for Agents," 18 de diciembre de 2024, https://www.anthropic.com/engineering/writing-tools-for-agents/.

[13] Unión Europea, *Reglamento (UE) 2024/1689 del Parlamento Europeo y del Consejo, de 13 de junio de 2024, por el que se establecen normas armonizadas en materia de inteligencia artificial (Artificial Intelligence Act)*, DO L, 2024/1689, 12 de julio de 2024, art. 14.

agéntica, el control efectivo es trazabilidad. Si el sistema actúa, debe quedar constancia de qué herramientas usó, con qué datos, bajo qué instrucciones, qué criterios activaron cada acción y qué incertidumbres estaban presentes. Sin trazabilidad, no hay aprendizaje (porque no se puede diagnosticar) y no hay responsabilidad (porque no se puede atribuir).

Una forma de operacionalizar esta idea es distinguir entre trazabilidad de evidencia (fuentes y datos), de decisión (prompts, reglas y umbrales) y de ejecución (acciones y llamadas a herramientas). Estas trazas conforman el expediente de una decisión híbrida y permiten auditoría, revisión y mejora; además, habilitan una supervisión selectiva, basada en señales, en lugar de una revisión exhaustiva e inviable.

Proponemos entender la gobernanza como un bloque de capacidades distribuida en cuatro niveles.

a) Nivel individual: alfabetización agéntica. Comprender qué es un agente, qué puede y qué no puede hacer, y cómo se degrada bajo incertidumbre, ambigüedad o inputs adversarios. Saber formular objetivos y exigir evidencia.

b) Nivel de equipo: protocolos compartidos. Plantillas de instrucciones, criterios de calidad, listas de verificación de seguridad, repertorio de evals, roles de verificación y revisión cruzada.

c) Nivel organizativo: infraestructura y políticas. Registro central de prompts y modelos, control de accesos, auditoría, gestión de incidentes, y un catálogo de herramientas con documentación diseñada para agentes. Aquí se decide qué tareas se automatizan, con qué grados de autonomía y qué indicadores de riesgo activan supervisión reforzada.

d) Nivel ecosistémico: interoperabilidad y estándares. Modelos de documentación (por ejemplo, para describir herramientas), prácticas de evaluación comparables y marcos regulatorios que armonicen expectativas.

Esta pila permite reencuadrar el papel de la regulación. El AI Act no define "competencias", pero al exigir supervisión humana en sistemas de alto riesgo, obliga a crear capacidades de intervención real: asignación de responsabilidad, formación adecuada, interfaz que permita interpretar el funcionamiento y mecanismos de parada. En otras palabras, obliga a que la supervisión sea un diseño, no una intención. En paralelo, marcos como el NIST AI RMF y su perfil para IA generativa ofrecen un vocabulario operativo para gestionar riesgos (gobernar, mapear, medir, gestionar) y pueden integrarse como prácticas organizativas[14].

El punto crucial es que gobernar no es frenar; es hacer que el sistema sea mejorable. En IA agéntica, los fallos no se eliminan por decreto, sino por iteración: se detectan, se clasifican, se convierten en evals, se ajustan instrucciones o herramientas, y se re-prueba. La gobernanza es el ciclo de aprendizaje institucional.

6. Formación y diseño organizativo para equipos con agentes

Si la IA agéntica se integra como "un empleado digital", la formación no puede limitarse a "cómo escribir prompts". Requiere un currículum de capacidades distribuido entre técnica, cognición y gobernanza. Proponemos cinco líneas de formación aplicables a todo tipo de organizaciones públicas o privadas.

1) Pensamiento de sistemas y descomposición de tareas. Aprender a convertir problemas difusos en flujos verificables, identificar dependencias, datos críticos y puntos de fallo. Esto incluye modelar el proceso antes de automatizarlo.

2) Verificación y evaluación. Entrenar en lectura crítica, contraste de fuentes, construcción de pruebas y uso de evals. La habilidad central es convertir un criterio de calidad en un test repetible. En redacción, listas de verificación; en datos, reproducibilidad; en software, tests; en decisiones, auditoría de supuestos.

3) Orquestación y herramientas. Conocer patrones de agentes (chaining, routing, orquestador-trabajadores) y aprender a diseñar herramientas ergonómicas para agentes, con entradas claras y salidas de alto valor informativo . Esto requiere pensar como diseñador de interfaces para un "usuario" no determinista.

[14] National Institute of Standards and Technology (NIST), *Artificial Intelligence Risk Management Framework (AI RMF 1.0)* (NIST AI 100-1) (Gaithersburg, MD: NIST, 2023), https://nvlpubs.nist.gov/nistpubs/ai/nist.ai.100-1.pdf.

4) Seguridad y privacidad operativa. Reconocer riesgos de inyección, filtraciones, uso indebido de datos y sobre-autonomía; aprender a usar sandboxes y entornos de prueba antes de producción; y entender que la privacidad es parte de la calidad del sistema.

5) Metacognición y aprendizaje deliberado. Diseñar rutinas personales y de equipo que mantengan habilidades internas. Por ejemplo, alternar ciclos de trabajo asistido y no asistido, y mantener "cuadernos de decisiones" donde se registren supuestos, evidencias y razonamientos.

Estas líneas pueden traducirse a programas formativos orientados a práctica, con casos de fallo, conjuntos de prueba y revisión de decisiones.

En paralelo, el diseño organizativo debe adaptarse. Equipos con agentes funcionan mejor cuando se redefinen responsabilidades: quién autoriza acciones, quién mantiene las evals, quién gestiona incidentes, quién decide cuándo actualizar modelos. También necesitan interfaces de confianza: visualizaciones de fuentes, trazas de herramientas y explicaciones estructuradas que permitan inspección. La transparencia no es un valor abstracto; es una condición para que el juicio humano sea posible.

Proponemos tres recomendaciones de diseño:

- Diseñar para la inspección: todo flujo agéntico debe producir un "paquete de evidencia" que permita revisión rápida (fuentes, decisiones, ejecución). Si no se puede inspeccionar, no se debe delegar con autonomía alta.
- Diseñar para el aprendizaje: cada fallo relevante debe convertirse en un caso de prueba. Esto crea una memoria organizativa y evita que el sistema repita errores con aparente seguridad.
- Diseñar para la reversibilidad: las acciones irreversibles (enviar, borrar, aprobar) deben requerir confirmación humana o mecanismos de doble control. La autonomía debe crecer solo donde la reversibilidad y la verificabilidad lo permitan.

7. Aplanamiento organizativo y equipos híbridos en la era de la IA agéntica

La irrupción de sistemas de IA cada vez más capaces y, en particular, de la IA agéntica que no solo "asiste" sino que planifica, encadena subtareas y ejecuta acciones, reconfigura la morfología de la organización. El efecto más visible es un aplanamiento estructural: disminuye la necesidad de capas intermedias dedicadas a coordinar, consolidar información, traducir decisiones en tareas y verificar entregables.

En paralelo, emerge la forma de trabajo dominante de la co-inteligencia: el equipo híbrido, entendido no como un grupo humano que usa herramientas, sino como una unidad operativa compuesta por personas y agentes con funciones diferenciadas, integradas por reglas de delegación, control y trazabilidad.

7.1. Por qué las organizaciones se aplanan: de la escasez de ejecución a la escasez de criterio

En los modelos organizativos tradicionales, buena parte de la jerarquía se justificaba por la escasez de capacidad de producción y por los costes de coordinación. El "mando" era, a la vez, mecanismo de asignación de tareas, control de calidad y canal de información. Cuando la IA agéntica abarata drásticamente la elaboración de borradores, el análisis preliminar, la síntesis documental, la generación de alternativas y, en algunos casos, la ejecución operativa, la organización experimenta un desplazamiento del cuello de botella: hacer deja de ser lo crítico; lo crítico pasa a ser decidir bien.

Este cambio tiene dos consecuencias. Primero, las capas intermedias centradas en supervisión micro-operativa pierden centralidad, porque la organización puede instrumentar controles automáticos (reglas, pruebas, trazas, auditorías) y verificación por muestreo en puntos de control. Segundo, la dirección se transforma: de una lógica de vigilancia del trabajo humano a una lógica de diseño del sistema socio-técnico (estándares, métricas, umbrales de riesgo, protocolos de escalado). En términos de gobernanza, el aplanamiento no implica ausencia de control, sino control reubicado: menos control "por presencia" y más control "por arquitectura".

7.2. Del organigrama al flujo decisional: autoridad distribuida y puntos de escalado

El aplanamiento organizativo no opera como una simple reducción de mandos, sino como una reingeniería del trabajo en torno a flujos decisionales. La autoridad se desplaza hacia los bordes de la organización en aquellas tareas que cumplen dos condiciones: (i) son repetitivas o suficientemente estandarizables y (ii) su perfil de riesgo permite delegación bajo reglas. En ese terreno, los equipos cercanos al problema pueden operar con mayor autonomía porque disponen de capacidades algorítmicas que antes estaban fragmentadas en funciones de soporte (análisis, documentación, control de calidad, reporting).

Sin embargo, el aplanamiento solo es sostenible si se institucionalizan mecanismos de escalado: reglas que determinan cuándo una decisión debe elevarse a revisión humana de mayor nivel o a un comité de riesgo. No se trata de crear burocracia, sino de fijar condiciones objetivas: impacto económico, afectación reputacional, incertidumbre, datos sensibles, cambios de política, desviación respecto de patrones previos. La organización co-inteligente se reconoce por este diseño: autonomía amplia en lo delegable y deliberación reforzada en lo crítico.

7.3. Equipos híbridos: de "herramientas" a "colegas operativos" con límites

El equipo híbrido es la respuesta funcional a la IA agéntica. En lugar de un equipo humano que "consulta" una IA, se consolida una configuración en la que distintos agentes desempeñan roles operativos relativamente estables (búsqueda y síntesis, generación de opciones, documentación, pruebas, monitorización), mientras que los humanos concentran el criterio, la priorización y la responsabilidad.

La hibridación introduce roles que, aunque pueden recaer en personas existentes, se vuelven explícitos:

- Propietario de decisión: asume la responsabilidad final y define los criterios de aceptabilidad (calidad, riesgos, trade-offs).
- Orquestador: descompone objetivos en subtareas, asigna agentes y establece límites; gestiona dependencias y "handoffs".
- Verificador: contrasta fuentes, valida cifras, detecta incoherencias y alucinaciones, revisa supuestos y trazabilidad.
- Arquitecto de calidad: convierte estándares en pruebas (rúbricas, checklists, tests de regresión, "definition of done"), y monitoriza deriva.

Estos roles no son decorativos: permiten que el aplanamiento no derive en descontrol. La organización co-inteligente no elimina la supervisión; la convierte en supervisión distribuida y formalizada mediante reglas y métricas.

7.4. Métricas y disciplina operativa: fiabilidad, trazabilidad y tiempo de verificación

En equipos híbridos, medir solo "output" es insuficiente y, de hecho, peligroso: incentiva velocidad sin garantías. Por ello, las organizaciones co-inteligentes introducen métricas de calidad estructural: tasa de errores detectados en verificación, densidad y calidad de evidencia, trazabilidad de decisiones (qué información se usó, qué agente intervino, qué controles se aplicaron), y tiempo de verificación relativo al tiempo de generación. La eficiencia relevante no es producir más, sino producir mejor con menos fricción de verificación.

Este giro métrico opera como condición de posibilidad del aplanamiento. Si la calidad está instrumentada, no hace falta interponer capas jerárquicas para garantizarla; basta con diseñar el sistema para que la calidad sea un "resultado por defecto".

7.5. Una tesis organizativa: menos jerarquía de vigilancia, más jerarquía de sentido

En síntesis, la organización co-inteligente se aplana porque la IA agéntica reduce costes de coordinación y multiplica capacidad de ejecución, pero se mantiene, e incluso se refuerza, una jerarquía distinta: la jerarquía del sentido y el criterio. La dirección ya no se legitima por controlar tareas, sino por fijar prioridades, diseñar umbrales, construir estándares y garantizar que la delegación algorítmica no erosione la calidad epistémica del trabajo. Los equipos híbridos, por

su parte, no son un "añadido tecnológico", sino la nueva unidad de producción: una cooperación sistemática entre capacidades humanas (juicio, responsabilidad, valores) y capacidades algorítmicas (escala, velocidad, exploración de alternativas), gobernada por reglas de delegación, verificación y aprendizaje institucional.

8. Conclusiones

La IA agéntica representa un cambio de fase en la integración de IA en el trabajo del conocimiento. El foco deja de ser la automatización de tareas aisladas y pasa a ser la delegación de flujos completos con capacidad de acción. Esto amplifica el potencial de productividad y nivelación de habilidades, pero también los riesgos de errores encadenados, inyección de instrucciones y sobre-autonomía. Para navegar esta transición, hemos propuesto tratar la co-inteligencia como un problema de capacidades: qué debe saber hacer una persona, qué debe institucionalizar un equipo y qué debe soportar una organización para que la delegación sea segura y útil. El marco de formulación, orquestación, verificación, calibración, resiliencia cognitiva, seguridad y gobernanza, ofrece un lenguaje operativo para diseñar prácticas y políticas. En este enfoque, la regulación europea sobre supervisión humana funciona como un marco habilitador: obliga a construir capacidades de intervención real y a sostener la trazabilidad que hace posible la rendición de cuentas.

La conclusión es que, cuanto más inteligente y autónoma sea la IA, más valiosas se vuelven las capacidades humanas de criterio, verificación y diseño de objetivos. No porque la máquina "no pueda" producir razonamientos, sino porque el valor social y organizativo del razonamiento depende de la responsabilidad, del aprendizaje y de la capacidad de dar razones. En la era de la IA agéntica, la agencia humana se preserva menos por proclamación y más por diseño.

Declaración sobre el uso de IA

Este manuscrito ha sido revisado y reescrito con apoyo de un sistema de IA generativa utilizado como herramienta de edición, reorganización y mejora estilística. La responsabilidad final del contenido, la selección de fuentes y la coherencia argumental corresponde a la autora.

Bibliografía

Anthropic. "Building Effective Agents." 19 de diciembre de 2024. https://www.anthropic.com/research/building-effective-agents.

"Engineering at Anthropic: Writing Tools for Agents." 18 de diciembre de 2024. https://www.anthropic.com/engineering/writing-tools-for-agents.

Brynjolfsson, Erik, Danielle Li y Lindsey Raymond. "Generative AI at Work." *The Quarterly Journal of Economics* 140, n.º 2 (2025): 889–942. https://doi.org/10.1093/qje/qjae044.

Dellermann, Dominik, Philipp Ebel, Matthias Söllner y Jan Marco Leimeister. "Hybrid Intelligence." *Business & Information Systems Engineering* 61, n.º 5 (2019): 637–643. https://doi.org/10.1007/s12599-019-00595-2.

National Institute of Standards and Technology (NIST). *Artificial Intelligence Risk Management Framework (AI RMF 1.0)*. NIST AI 100-1. Gaithersburg, MD: National Institute of Standards and Technology, 2023. https://nvlpubs.nist.gov/nistpubs/ai/nist.ai.100-1.pdf.

Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile. NIST AI 600-1. Gaithersburg, MD: National Institute of Standards and Technology, 2024. https://doi.org/10.6028/NIST.AI.600-1.

OpenAI. "A Practical Guide to Building Agents." Documento PDF. Consultado el 31 de diciembre de 2025. https://cdn.openai.com/business-guides-and-resources/a-practical-guide-to-building-agents.pdf.

"New Tools for Building Agents." 11 de marzo de 2025. https://openai.com/index/new-tools-for-building-agents/.

OWASP Foundation. "OWASP Top 10 for Large Language Model Applications (Version 1.1)." Consultado el 31 de diciembre de 2025. https://owasp.org/www-project-top-10-for-large-language-model-applications/.

Risko, Evan F., y Sam J. Gilbert. "Cognitive Offloading." *Trends in Cognitive Sciences* 20, n.º 9 (2016): 676–688. https://doi.org/10.1016/j.tics.2016.07.002.

Sparrow, Betsy, Jenny Liu y Daniel M. Wegner. "Google Effects on Memory: Cognitive Consequences of Having Information at Our Fingertips." *Science* 333, n.º 6043 (2011): 776–778. https://doi.org/10.1126/science.1207745.

Sperber, Dan, Hugo Mercier, Christophe Origgi y otros. "Epistemic Vigilance." *Mind & Language* 25, n.º 4 (2010): 359–393. https://doi.org/10.1111/j.1468-0017.2010.01394.x.

Unión Europea. *Reglamento (UE) 2024/1689 del Parlamento Europeo y del Consejo, de 13 de junio de 2024, por el que se establecen normas armonizadas en materia de inteligencia artificial (Artificial Intelligence Act)*. DO L, 2024/1689, 12 de julio de 2024. https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng