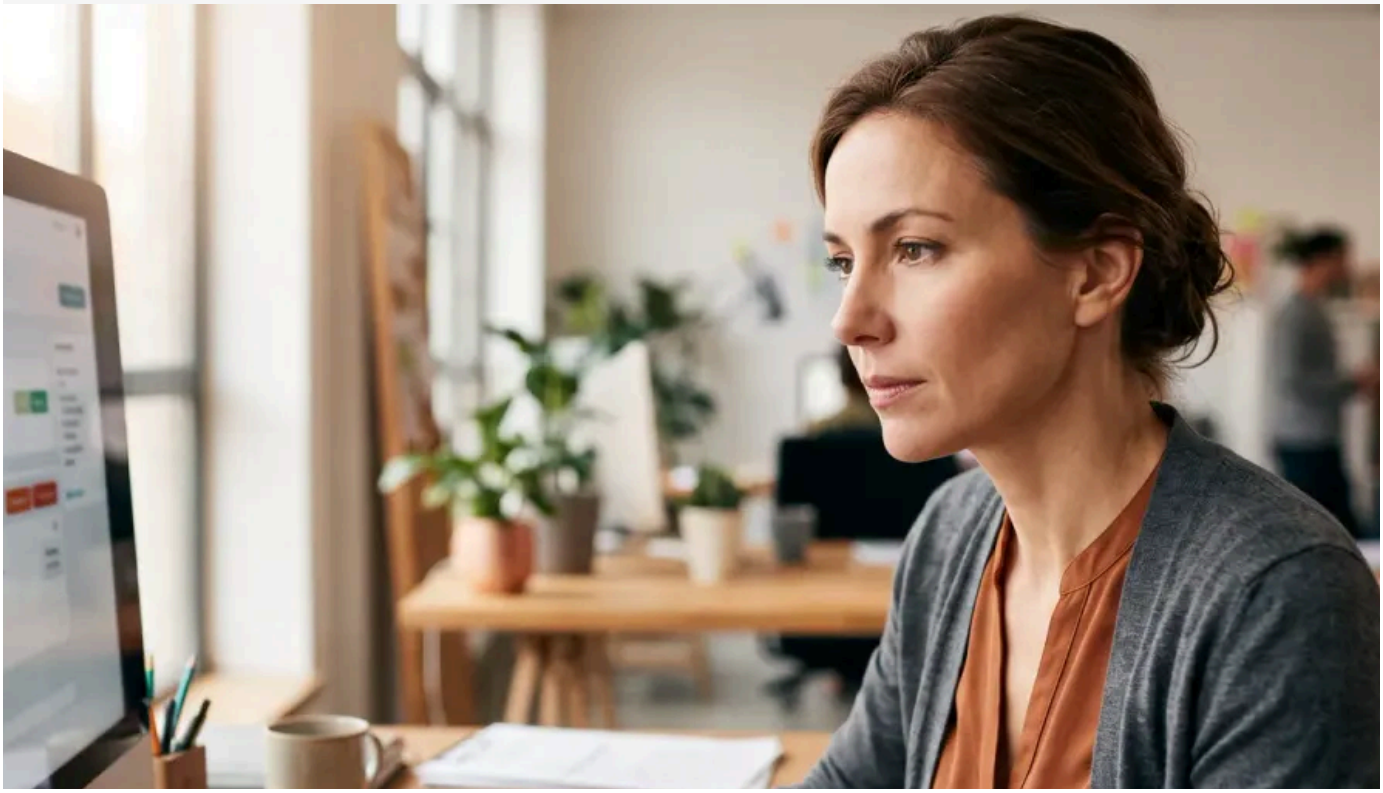
Agentes de IA con control humano: dónde poner las validaciones

Un agente de IA sin supervisión es un riesgo y un agente con supervisión en todo es un coste. La pregunta útil no es si poner a una persona en el circuito, sino dónde. Esta guía te da cuatro niveles de autonomía, una herramienta para elegir el tuyo y diez controles que usamos en proyectos con clientes.

Por **Valentín Ayesa**, fundador de VA360 Labs · 13 min de lectura · Revisado el 5 de noviembre de 2026





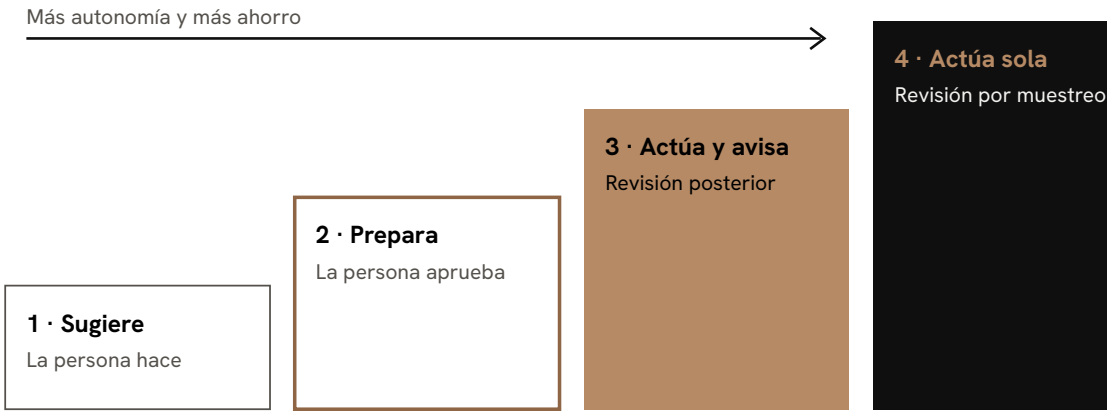
La respuesta en un minuto

Pon a una persona **donde un error sea difícil de deshacer o afecte a un cliente, a dinero o a datos personales**, y deja que el agente trabaje solo en lo demás. La autonomía no es un interruptor: son cuatro niveles, y cada tarea de un mismo agente puede estar en uno distinto.

Un buen diseño no hace que la persona rehaga el trabajo: le deja *validar* un resultado con criterios claros.

Cuatro niveles de autonomía

Casi todas las discusiones sobre «control humano» se resuelven al ponerle nombre al nivel. Estos son los cuatro que usamos al diseñar un agente:



A más autonomía, más ahorro y más exposición. El nivel correcto depende de lo que cuesta un error, no de lo bueno que sea el modelo.

Qué significa cada nivel en la práctica

Nivel	Ejemplo	Dónde está la persona
1 · Sugiere	Propone una respuesta a una reclamación	Escribe y envía ella
2 · Prepara y espera	Deja el reembolso preparado con el importe y el motivo	Aprueba antes de que se ejecute
3 · Actúa y avisa	Clasifica y etiqueta tickets, y avisa de lo que no encaja	Revisa a posteriori y puede deshacer
4 · Actúa sola	Extrae datos de documentos de formato fijo	Revisa una muestra y vigila las métricas

Versión impresa: abre la guía en va360labs.com para usar la herramienta con tus datos.

Elige el nivel de tu agente

Responde sobre lo que hace tu agente. La recomendación y los controles asociados se actualizan al momento. Se calcula en tu navegador y no se guarda nada.

¿El resultado llega a un cliente o a un tercero sin que nadie lo vea antes?

Sí | **No**

¿Mueve dinero, cambia precios o concede algo (reembolsos, descuentos, condiciones)?

Sí | **No**

¿Trata datos personales o información confidencial?

Sí | **No**

¿Un error es difícil o caro de deshacer?

Sí | **No**

¿Una norma o un contrato exige decisión humana o trazabilidad?

Sí | **No**

¿Los casos son muy variados o a menudo ambiguos?

Sí | **No**

¿El volumen es tan alto que revisarlo todo es inviable?

Sí | **No**

Nivel recomendado: 4

La IA actúa sola con revisión por muestreo.

1 **La IA sugiere, la persona decide y ejecuta**

La IA propone y no toca nada. Es el nivel de entrada cuando el riesgo

2 **La IA prepara todo y espera aprobación**
Deja el trabajo hecho hasta el último paso y una persona aprueba antes de que tenga efecto.

3 **La IA actúa en lo reversible y avisa**
Ejecuta lo que se puede deshacer y una persona revisa después, con capacidad de corregir.

4 **La IA actúa sola con revisión por muestreo**
Solo para tareas de bajo riesgo y mucho volumen, con métricas y alertas.

Controles que te recomendamos para este caso

- [ ] Un responsable con nombre y apellido para este agente.
- [ ] Un registro de todo lo que hace, consultable por quien lo supervisa.
- [ ] Un interruptor de parada que cualquiera del equipo pueda accionar.



La validación útil no es un clic automático: es alguien con criterio que sabe qué mirar.

Lo que enseñan los proyectos reales

En los proyectos de agentes que hemos entregado, la persona no está en el mismo sitio. Esto es lo que dicen los casos publicados:

Dónde está la persona en cinco proyectos publicados

Proyecto	Lo que hace la IA	Dónde interviene una persona
Avatar TC (The Clueless)	Genera caras, variaciones, modelos y contenido	Cinco puertas de validación por rol (cara, variaciones, LoRA, pack web y contenido), con registro de quién validó y cuándo
Agente de soporte (MyTopTour)	Clasifica tickets y consulta reservas en HubSpot y Salesforce	Revisión en cambios, cancelaciones y quejas; los datos personales se enmascaran antes de llegar al modelo
Wayla	Responde el correo de soporte de empresas de viajes	Deriva a una persona cuando no hay una respuesta fiable; ningún agente se publica sin simulaciones aprobadas y la firma de una persona
LinkedFlow	Ejecuta flujos de prospección y personaliza mensajes	Límites, ventanas horarias, espaciado entre acciones y cortacircuitos; la automatización se detiene sola cuando el prospecto responde
Agente de voz (Cabify Empresas)	Llama, responde objeciones, cualifica y agenda	La reunión posterior la sigue haciendo una persona del equipo comercial

Hay un patrón: **la persona aparece en el punto de decisión o de salida, no en el trabajo intermedio**. El agente hace lo laborioso y la persona valida lo que importa. Por eso en Avatar TC el tiempo humano baja de unas 14 horas por avatar a unas 3 según nuestra estimación, aunque hay cinco validaciones por medio.

Diez controles antes de publicar

Esta es la lista que repasamos antes de poner un agente delante de clientes. No es exhaustiva, pero cada punto ha evitado un problema real.

01

Un responsable con nombre

Una persona concreta responde de lo que hace el agente. «El equipo» no es un responsable.

02

Casos de prueba reales antes de activar

Simula al menos cincuenta casos con datos de verdad, incluidos los difíciles. Publica solo con resultados aprobados y firmados.

03

Un camino claro hacia una persona

El cliente debe poder hablar con alguien sin pelearse con el agente. Y el agente debe saber cuándo derivar.

04

Umbral de confianza

Si no hay una respuesta fiable, deriva. Un «no lo sé» es mejor que una invención.

05

Límites duros

Importe máximo, número de acciones por hora y ventanas horarias. Los límites los impone el sistema, no se piden en el texto de instrucciones.

06

Datos personales protegidos

Enmascara lo que no necesite el modelo, define cuánto se conserva y quién puede verlo.

07

Registro de todo

Qué recibió, qué decidió, quién lo aprobó y con qué versión del modelo y de las instrucciones.

08

Interruptor de parada

Que cualquier persona del equipo pueda detener el agente en un clic, sin pedir permiso a desarrollo.

09

Prueba de regresión cuando algo cambia

Un cambio de modelo, de instrucciones o de datos puede romper lo que funcionaba. Repite los casos de prueba antes de desplegar.

10

Revisión periódica

Al principio, revisión semanal de una muestra de conversaciones. Después, mensual: errores, derivaciones y quejas.

Qué dice la normativa

La normativa no te obliga a poner a una persona en todo, pero sí en algunos puntos concretos. Resumimos lo que más nos preguntan:

- **Reglamento europeo de IA (UE 2024/1689).** Para los sistemas de alto riesgo, el artículo 14 exige que se diseñen para poder ser vigilados eficazmente por personas. Y el artículo 50 obliga a informar a quien interactúa con un sistema de IA, salvo que sea evidente.
- **RGPD, artículo 22.** Las personas tienen derecho a no ser objeto de una decisión basada únicamente en un tratamiento automatizado que produzca efectos jurídicos o les afecte de modo significativo. Si el agente decide sobre créditos, empleo o reclamaciones, necesitas una persona que intervenga.

> **Esto no es asesoramiento jurídico.** Si tu agente toca decisiones sobre personas, dinero o salud, consulta con tu asesoría legal o con tu delegado de protección de datos antes de publicarlo.

Preguntas frecuentes

¿Qué es un agente de IA con control humano?

Es un agente que prepara o ejecuta tareas pero cuyas decisiones de más riesgo pasan por una persona antes de producir efectos: enviar un mensaje a un cliente, mover dinero, cambiar un dato maestro. Se suele llamar human in the loop (persona en el circuito).

¿Dónde tiene que haber siempre una persona?

Donde el error sea difícil de deshacer o afecte a un cliente, a dinero o a datos personales: cambios, cancelaciones, reclamaciones, pagos y comunicaciones

nuevas con un cliente. En esos puntos el agente prepara y una persona aprueba.

¿La supervisión humana elimina el ahorro?

No si se diseña bien. En el proyecto de Avatar TC, la plataforma exige cinco validaciones humanas por avatar y aun así reduce el tiempo humano por avatar de unas 14 horas a unas 3 según nuestra estimación. La clave es que la persona valide resultados con criterios claros, no que rehaga el trabajo.

¿Hay que informar al cliente de que habla con una IA?

Es una buena práctica y, según el caso, una obligación. El Reglamento europeo de IA exige informar a las personas cuando interactúan con un sistema de IA, salvo que sea evidente. Recomendamos hacerlo siempre y dejar una vía clara para hablar con una persona.

¿Cada cuánto hay que revisar un agente en producción?

Al principio, a diario o por muestreo semanal de conversaciones; después, con una revisión mensual de errores y escaladas y una prueba de regresión cada vez que cambia el modelo, el prompt o los datos. Un agente que funcionó ayer puede fallar hoy si cambia el modelo que lo sostiene.

Fuentes y transparencia

- **Reglamento (UE) 2024/1689 de Inteligencia Artificial, artículos 14 (supervisión humana) y 50 (transparencia)**
- **VA360 Labs. Caso Avatar TC (The Clueless): cinco puertas de validación humana**
- **VA360 Labs. Caso MyTopTour: agente de soporte con revisión humana en cambios, cancelaciones y quejas**

- Las fotografías de esta guía están generadas con IA y no muestran personas ni empresas reales. Los diagramas y la herramienta son de VA360 Labs.
- Las cifras de nuestros proyectos proceden de los informes de entrega y se explican en cada caso. Los ejemplos con números que no son de un caso se indican como ilustrativos.



Valentín Ayesa

Embajador oficial de n8n en España, Meta Tech Partner y Claude Partner. Dirige VA360 Labs, donde diseña y construye agentes y automatizaciones para empresas, y firma las guías de esta web. Más sobre Valentín.

Casos donde lo aplicamos

The Clueless

Una plataforma para producir influencers virtuales en serie, con validación humana en cada fase

Un proceso que se hacía avatar a avatar y a mano pasa a una única plataforma con cinco puertas de validación humana, pensada para un plan de 1.000 avatares.

MyTopTour

Un agente de soporte con IA en los servidores del cliente, conectado a HubSpot y Salesforce

En la prueba de integración con Salesforce identificó 49 de 60 tickets en 15 minutos, sin errores.

Wayla

Wayla: agentes de IA que atienden el correo de soporte de las empresas de viajes

Ningún agente se publica sin simulaciones aprobadas y la firma de una persona.

Más guías

14 min

Calcular el retorno

Método para calcular el retorno de un proyecto de IA con la fórmula que usamos en VA360 Labs, errores habituales y una calculadora que funciona en tu navegador.

12 min

Qué automatizar primero

Cómo elegir qué proceso automatizar primero con IA: cinco criterios puntuables, un veto de riesgo y una herramienta que ordena tus procesos en tu navegador.

15 min

MCP para empresas

Qué es el Model Context Protocol (MCP), en qué se diferencia de una API o de n8n y cómo conectar una IA a los sistemas de tu empresa con permisos mínimos.

Servicio

Desarrollo con IA

Cómo lo hacemos con empresas medianas y grandes.

¿Lo aplicamos a tu empresa?

Esta guía es una versión de https://va360labs.com/guias/agentes-ia-control-humano/. La herramienta interactiva, las actualizaciones y el resto de guías están en la web.

Escríbenos a **valen@va360labs.com** o pide una primera conversación en va360labs.com. VA360 Labs S.L. · CIF B24805335 · Arcos de la Frontera (Cádiz).