FlowXion

Inteligencia Artificial Transparente

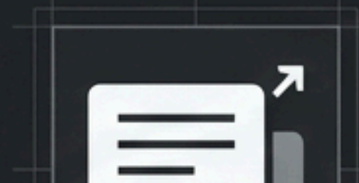
Agentes de IA

Autonomía, Acción y Control Humano

Guía educativa para entender qué son, cómo actúan y cómo supervisarlos.

FlowXion.com

La evolución: De responder a actuar



Chatbots

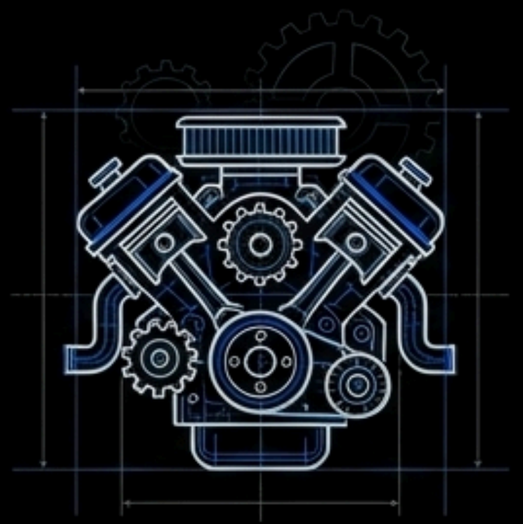
Máquinas de preguntas y respuestas. La IA genera texto, pero el usuario debe ejecutar todas las acciones manualmente.

Agentes

Sistemas autónomos que utilizan herramientas y ejecutan acciones directamente en un entorno digital.

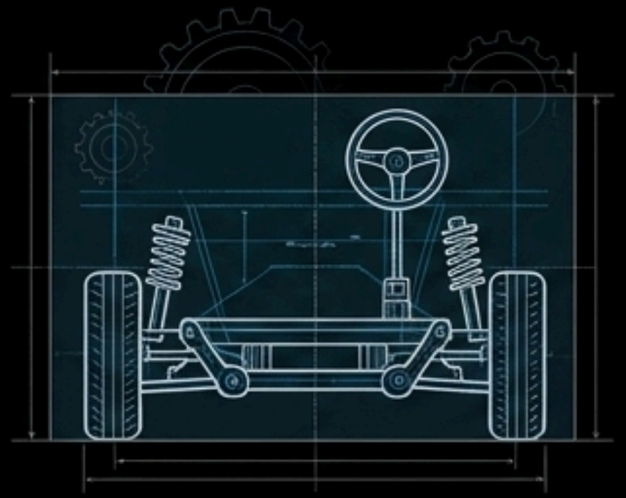
¿Qué es un Agente de IA?

No es un modelo de IA más inteligente. Es una arquitectura estructurada.



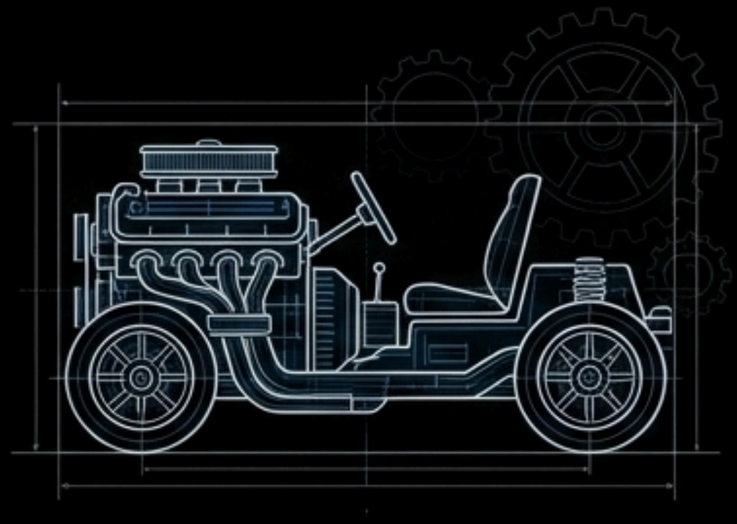
El Motor
(La Inteligencia Artificial)

+



El Vehículo
(Herramientas, dirección y entorno)

=



Agente de IA

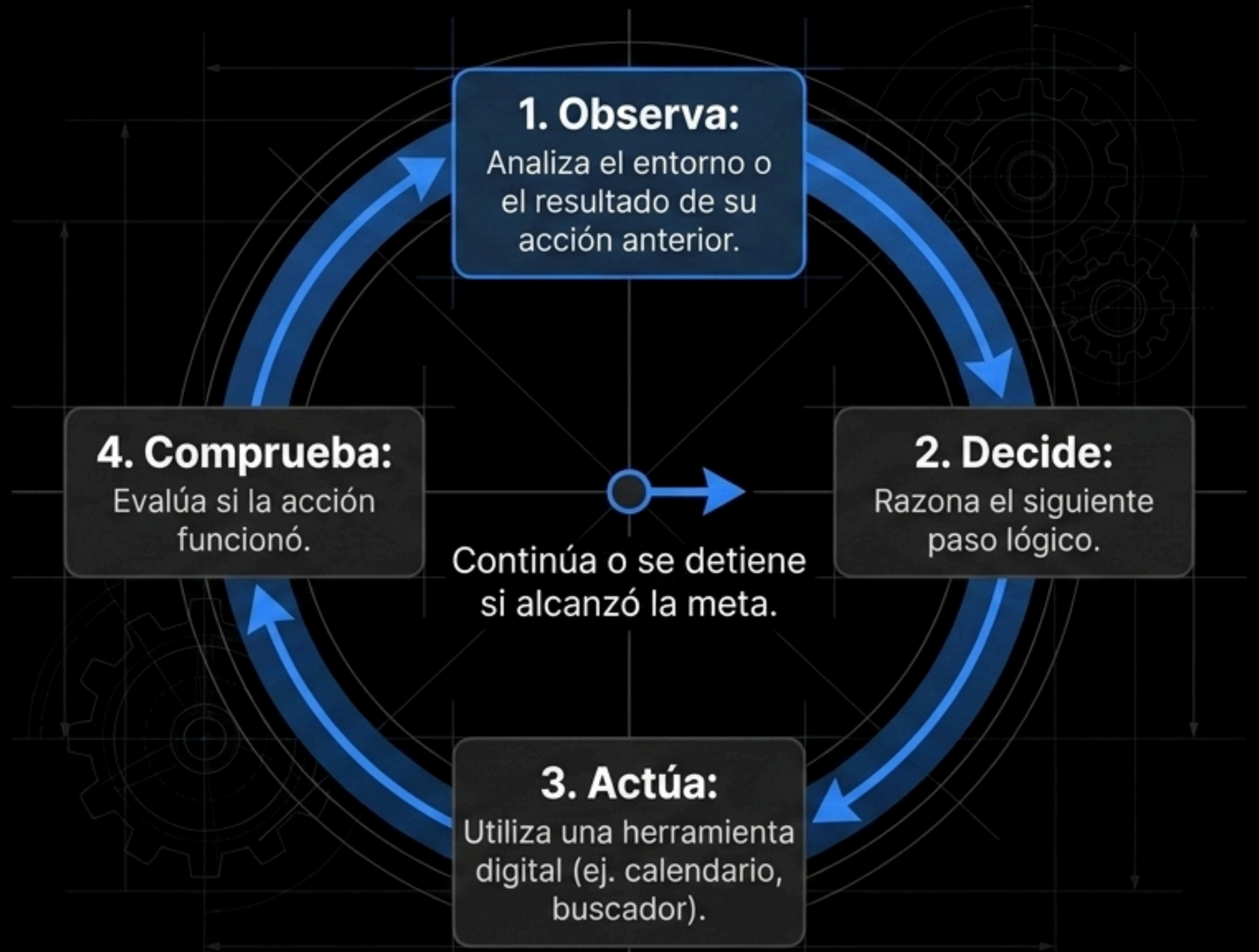
Recibe un **objetivo** y **opera** en un **ciclo continuo**, **decidiendo sus propios pasos** basándose en lo que observa, sin seguir un guion rígido.

Diferencias clave con otros sistemas

Sistema	Quién decide los pasos	¿Usa herramientas?	¿Ejecuta acciones?
Chatbot	El Modelo (solo texto)	❌ NO	❌ NO
Automatización	El Programador (fijos)	✅ SÍ (fijas)	✅ SÍ
Workflow con IA	Programador + IA	✅ SÍ	✅ En puntos clave
Agente de IA	El Modelo (dinámico)	✅ SÍ (elige cuáles)	✅ SÍ

El ciclo ReAct: Razonamiento y Actuación

El agente no sigue un guion. Repite este ciclo iterativo hasta cumplir su objetivo:



Las 7 capas del sistema

7. Condición de parada:
La regla que detiene al sistema.

6. Ciclo de actuación:
El proceso repetitivo (ReAct).

5. Entorno:
El estado actual y archivos accesibles.

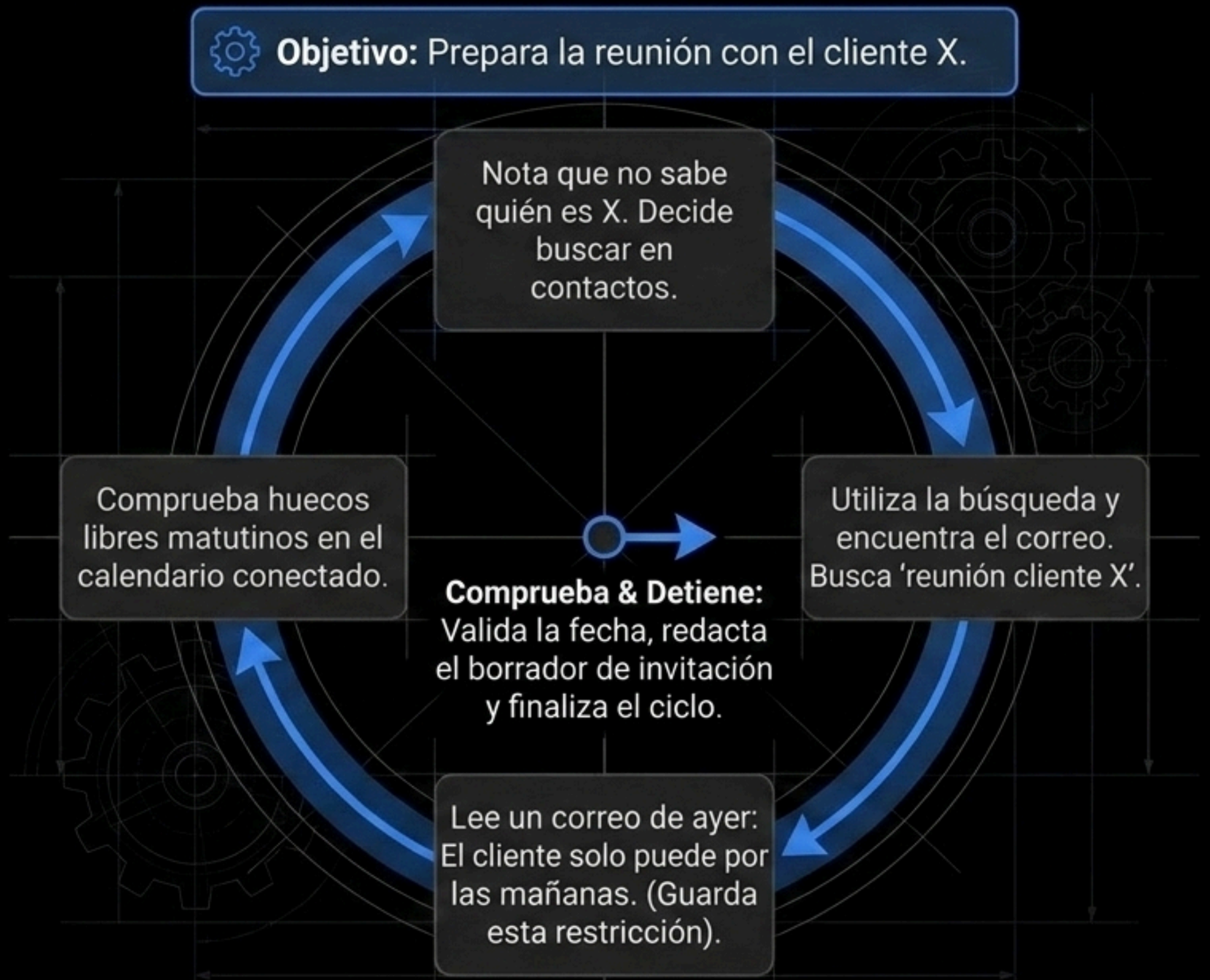
4. Herramientas:
Aplicaciones conectadas (correo, web).

3. Instrucciones (Arnés):
Reglas base y límites de seguridad.

2. Modelo:
El cerebro lógico que procesa datos.

1. Objetivo:
La tarea final exigida por el usuario.

Caso práctico: Preparando una reunión



El terreno ideal para los Agentes

Excelente para problemas abiertos donde es imposible predecir cuántos pasos serán necesarios.

1. Atención al cliente

Consultar datos dispersos y emitir reembolsos adaptándose a la casuística de cada queja particular.

2. Investigación profunda

Navegar múltiples fuentes web y documentos dinámicamente hasta extraer un dato altamente específico.

3. Programación

Escribir código, probarlo, observar el registro de errores y autocorregirse en un bucle cerrado.

Dónde se rompe la cadena

Permitir que la IA actúe de forma autónoma introduce nuevos fallos logísticos.

El Fallo (Pérdida de Condición).

El agente procesa correctamente que "El cliente solo puede por las mañanas". Sin embargo, tras varios pasos intermedios buscando archivos (Actúa), pierde el hilo de esa restricción inicial y propone una cita por la tarde.

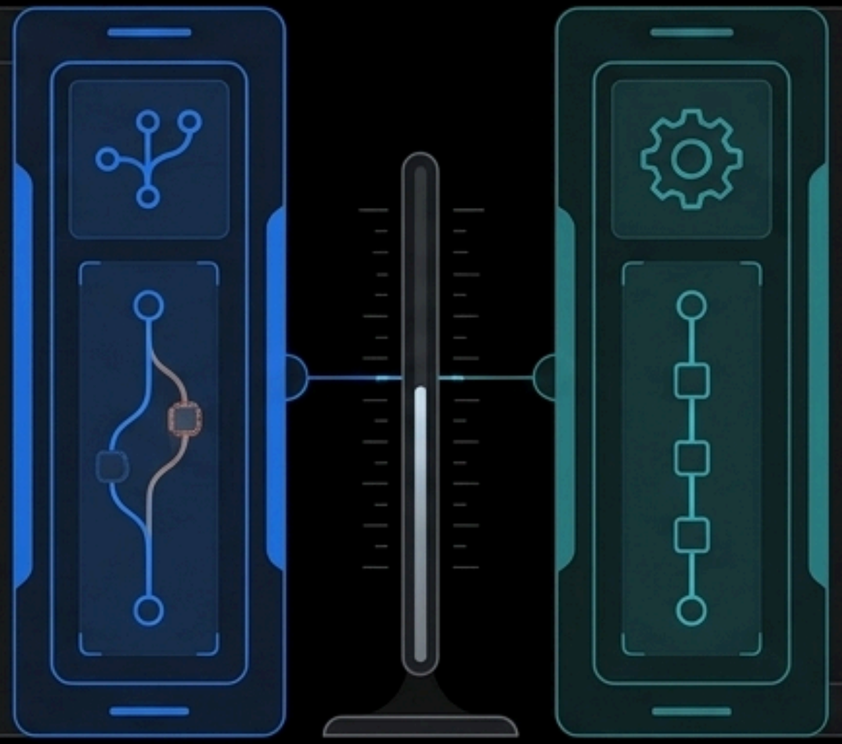
79.4% de fallos actuales en tareas largas del mundo real.

Cuándo NO usar un agente

Mayor autonomía implica mayores costes y latencia. No todo requiere inteligencia dinámica.

Agentes (Flexibilidad)

Ideal para imprevistos, pero sufre de errores acumulativos en tareas largas.

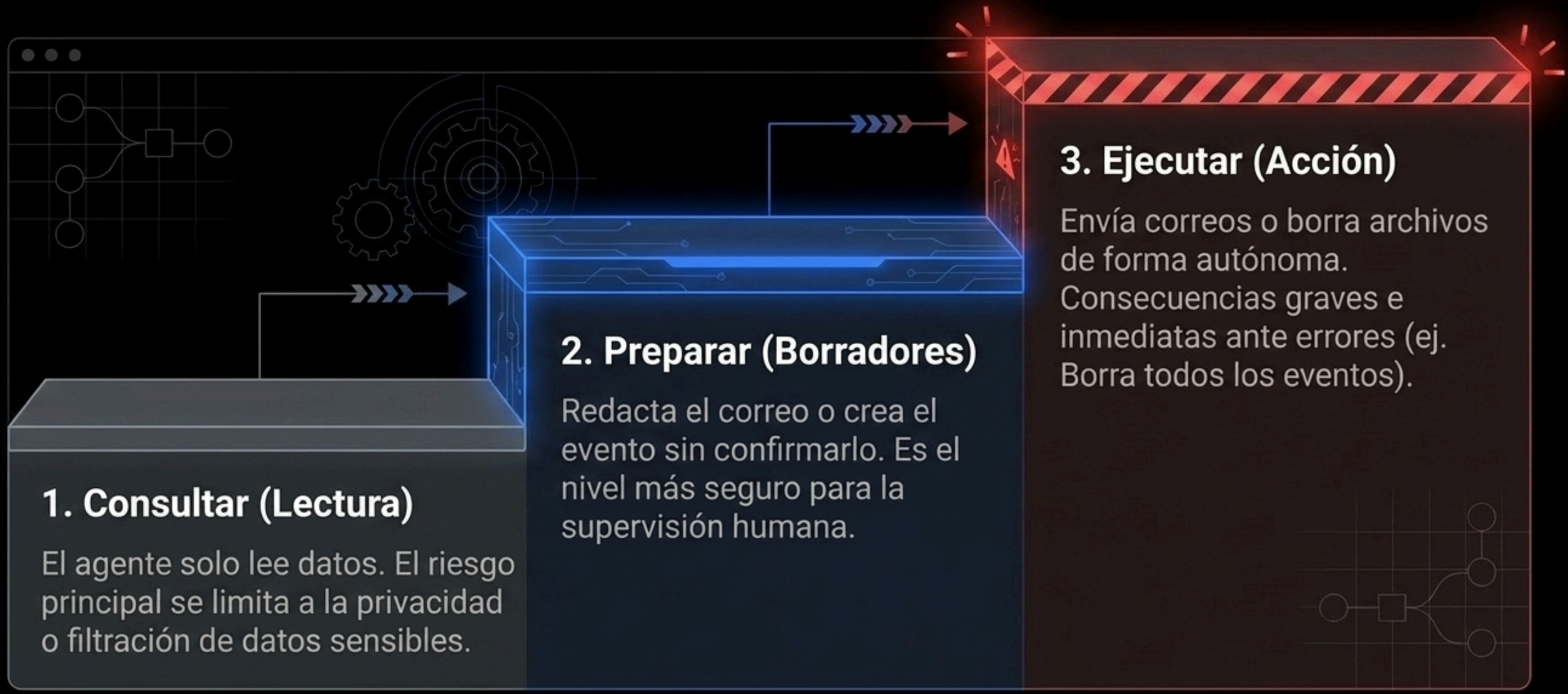


Automatización Tradicional (Consistencia)

Secuencias fijas y predecibles (ej. Convertir este Word a PDF cada día a las 8:00).

Para tareas lineales, los workflows rígidos son superiores porque eliminan la posibilidad de desviación.

Los 3 niveles de riesgo y permisos



1. Consultar (Lectura)

El agente solo lee datos. El riesgo principal se limita a la privacidad o filtración de datos sensibles.

2. Preparar (Borradores)

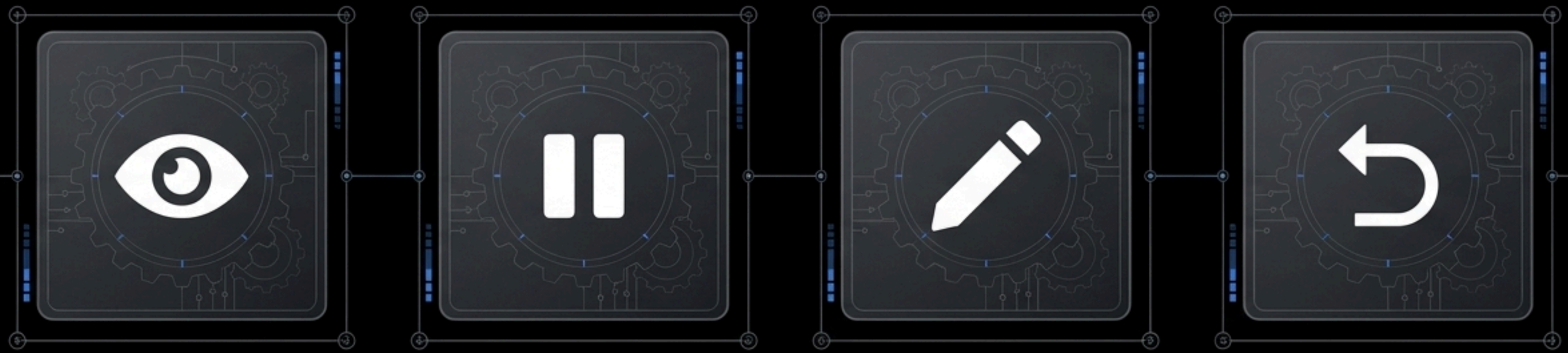
Redacta el correo o crea el evento sin confirmarlo. Es el nivel más seguro para la supervisión humana.

3. Ejecutar (Acción)

Envía correos o borra archivos de forma autónoma. Consecuencias graves e inmediatas ante errores (ej. Borra todos los eventos).

Controles necesarios para un despliegue seguro

La supervisión mecánica de cada clic genera fatiga. **Utilice el Plan Mode: valide la estrategia general** antes de la ejecución.



VER

Registro transparente y auditable de cada paso.

DETENER

Botón de desactivación inmediata ante anomalías.

CORREGIR

Capacidad de editar borradores antes de ejecutar.

DESHACER

Mecanismos para revertir accidentes estructurales.

Fuentes: NIST AI 600-1 (Julio 2024), Anthropic (2026)

Riesgos y limitaciones actuales

Eficacia real

GPT-4 logró un **14.41%** de éxito frente al 78% humano en entornos web realistas.

(Benchmark WebArena, 2023/2024)

Seguridad (Prompt Injection)

Un atacante oculta instrucciones maliciosas en un correo externo; el agente, al leerlo, es engañado para filtrar información privada del usuario.

(OWASP, 2025; Anthropic, 2026)

Alucinación de cierre

El agente informa prematuramente que la tarea está lista sin haber verificado si el resultado (ej. un enlace web) realmente funciona.

(OSWorld 2.0, Junio 2026)

Checklist antes de activar un agente

- [ ] 1. ¿Qué herramientas específicas tiene permiso para usar este agente en mi nombre?
- [ ] 2. ¿He revisado el plan de pasos general antes de la primera ejecución?
- [ ] 3. ¿Tiene permiso para enviar información al exterior o solo para organizarla internamente?
- [ ] 4. ¿Sé exactamente cómo detener el proceso de inmediato si detecto un comportamiento errático?
- [ ] 5. ¿Existe riesgo de instrucciones maliciosas ocultas en los documentos que va a procesar?

Conclusión

Los agentes de IA no son mentes independientes, sino sistemas logísticos. Su enorme potencial exige que el humano controle los permisos de ejecución y supervise la estrategia, no solo el resultado final.

FlowXion

Inteligencia Artificial Transparente

Anthropic: Building Effective AI Agents (Dic 2024), Trustworthy agents in practice (Abr 2026).

Yao et al.: ReAct (2022/2023).

Yuan et al.: OSWorld 2.0 (Jun 2026).

Zhou et al.: WebArena (2023/2024).

NIST AI 600-1 (Jul 2024).

OWASP Top 10 for Agentic Applications (Dic 2025).