

L U M B R E



Supervisar al agente: límites, permisos y verificación

Materia: Desarrollo de Software con IA

07/10/2026

Antes de empezar

Un agente con manos propias puede instalar dependencias, borrar archivos o ejecutar comandos que nunca autorizaste. La supervisión profesional no consiste en mirar la pantalla con miedo: consiste en configurar límites antes de empezar —alcance, aprobaciones, prohibiciones— y verificar después con evidencia: diffs, pruebas, comandos ejecutados. Este OA te da el reglamento completo que se le pega al documento maestro para que el agente trabaje solo... dentro de tu jaula.

QUÉ APRENDERÁS

- Dar alcance, aprobaciones y criterios a un agente que toca archivos y ejecuta comandos.

Autonomía sin jaula es un incidente esperando

Le diste al agente una tarea de tres archivos y terminó tocando quince, instalando un paquete que nadie pidió y "arreglando" un test que estorbaba. No es maldad: es la naturaleza de un sistema optimizado para cumplir tu objetivo con lo que ve. La diferencia entre el agente que usas y el agente que te usa está en el **reglamento**: qué puede tocar, qué debe pedirte, qué está prohibido y cómo te demuestra que cumplió.

ACTIVIDAD INTERACTIVA DISPONIBLE EN
LUMBRE

1. Las cuatro capas de la jaula



El reglamento, línea por línea

Nota que la prohibición de tocar pruebas es la más violada en la práctica: al agente "le conviene" que el test pase, y el camino más corto a veces es mover la meta.

Capa	Se le dice al agente	Por qué está
Alcance	«esta tarea toca solo src/forms y tests; si necesitas otro archivo, detente y dime por qué»	el agente no sabe qué es "suficiente"; el alcance le da frontera
Aprobaciones	«antes de instalar paquetes, borrar archivos o ejecutar algo contra producción, párate y pregúntame»	hay acciones con efectos que sobreviven a la sesión

Capa	Se le dice al agente	Por qué está
Prohibiciones	«nunca leas .env ni muestres claves; nunca modifiques una prueba para que pase; nunca hagas commit de lo que no pedí»	las líneas rojas no se negocian a mitad de tarea
Evidencia	«al terminar: lista de archivos tocados con diff, comandos ejecutados, pruebas corridas con resultado»	el reporte es el contrato: sin evidencia, el trabajo no está hecho

El bloque que va en tu documento maestro

El reglamento vive en el repo, no en tu memoria

```
## Reglamento para agentes (aplica a toda tarea)\n\nALCANCE: cada tarea declara los archivos que puede tocar. Si necesitas uno fuera,\ndetente y pide autorizacion con motivo.\n\nAPROBACION REQUERIDA (nunca ejecutes sin preguntar):\n- instalar o actualizar dependencias\n- borrar o mover archivos, cambiar la estructura de carpetas\n- cualquier comando contra produccion o con dinero\n- git commit / push\n\nPROHIBIDO SIEMPRE:\n- leer, imprimir o copiar .env, claves o tokens\n- modificar una prueba para que pase (si la prueba estorba, reporta el conflicto, no la editas)\n- "limpiar" codigo que no te pedi (no reformatees, no renombres)\n\nAL TERMINAR, REPORTA:\n1. archivos tocados + diff resumido  2. comandos ejecutados con su salida\n3. pruebas corridas y resultado real  4. lo que NO hice y por que\n\nYo reviso el reporte antes de dar la tarea por cerrada.
```

Para reforzar

Prueba

2. El agente reporta: «terminé; pasé las 12 pruebas». ¿Qué es lo mínimo que debes verificar antes de cerrar la tarea?

- ☐ Nada: el agente corrió las pruebas y dijo verde.
- ☐ Que el reporte incluya los archivos tocados y que entre ellos no esté el archivo de la prueba que fallaba ayer.
- ☐ Pedirle que corra de nuevo las pruebas frente a ti.
- ☐ Hacer commit inmediato antes de que se le olvide.

3. Pedir aprobación para instalar dependencias le quita velocidad al agente, así que es mejor desactivarlo.

☐ Verdadero

☐ Falso

Detecta la supervisión de confianza

ACTIVIDAD INTERACTIVA
DISPONIBLE EN LUMBRE

**4. Un alumno configura a su agente.
Señala la decisión que garantiza el incidente:**



5. Escribe hoy el bloque de reglamento en tu documento maestro (o créalo si aún no existe) con tus cuatro capas adaptadas a tu proyecto. En tu próxima tarea de agente, verifica el cierre contra el punto 4 de la evidencia: ¿te reportó también lo que NO hizo?

Supervisar no es vigilar con ansiedad: es diseñar la jaula una vez y caminar tranquilo.

Referencia curricular

Materia:	Desarrollo de Software con IA
-----------------	-------------------------------

Versión interactiva completa (con actividades, videos y autoevaluación):



<https://empresa.lumbre.lat/object/fsi-m19-0a02>

Material educativo generado desde Lumbre · lumbre.lat