

CÓDIGO	AUTOR	DURACIÓN ESTIMADA	NIVEL DE FORMACIÓN
01B06C07	Carlos Ojeda Sánchez	25 h	Avanzado

Dirigido a

Responsables de IA, seguridad, cumplimiento, protección de datos, transformación digital y áreas de negocio que necesiten implantar mecanismos de supervisión y control sobre sistemas agénticos.

Descripción

Con este contenido de curso profesional el alumnado aprenderá a establecer un marco de gobernanza para controlar agentes con distintos niveles de autonomía. Se priorizarán la monitorización, los guardrails, la protección de datos, la contención de fallos, el escalado a personas y la auditoría. La práctica permitirá integrar estos elementos en un plan de gobernanza proporcionado al riesgo del sistema.

COMPETENCIAS

1. Diseñar mecanismos de supervisión y monitorización para agentes autónomos, definiendo métricas, alertas y niveles de intervención.
2. Implementar guardrails de seguridad en agentes, diseñando límites, validaciones y respuestas controladas para reducir riesgos operativos.
3. Aplicar medidas de protección de datos y contención de fallos en agentes para reducir exposición, errores y usos indebidos.
4. Diseñar protocolos de escalado humano, definiendo patrones y puntos de intervención para controlar decisiones sensibles o inciertas.
5. Aplicar el marco ACPC y requisitos de auditoría para garantizar control, trazabilidad y revisión de sistemas agénticos.

CRITERIOS DE EVALUACIÓN (Objetivos)

1. Diferenciar niveles de supervisión y explicar cuándo conviene aplicar revisión previa, seguimiento continuo o control posterior. Definir métricas de rendimiento, calidad, seguridad y uso que permitan evaluar el comportamiento del agente. Diseñar alertas y umbrales de intervención asociados a errores, desviaciones, riesgos o resultados de baja confianza. Proponer un plan de monitorización que combine autonomía operativa con revisión humana proporcionada al nivel de riesgo.
2. Diferenciar tipos de guardrails y explicar su función preventiva, correctiva o limitadora en sistemas agénticos. Diseñar reglas de validación y bloqueo para controlar entradas, acciones, salidas y uso de herramientas sensibles. Aplicar criterios de seguridad y proporcionalidad para seleccionar guardrails adecuados al riesgo del caso de uso.
3. Especificar medidas básicas de protección de datos aplicables al diseño y uso de agentes en entornos profesionales. Explicar el riesgo de prompt injection y proponer medidas de mitigación proporcionadas al contexto de uso.

4. Diferenciar patrones de escalado humano y relacionarlos con niveles de riesgo, incertidumbre o impacto. Diseñar un protocolo de escalado completo que indique activadores, responsables, tiempos, evidencias y acciones posteriores.
5. Explicar los componentes del marco ACPC y su función dentro de una arquitectura de control agéntico. Especificar evidencias, registros y trazas necesarios para auditar decisiones, acciones, errores e intervenciones del agente. Proponer revisiones periódicas que permitan mantener controles actualizados ante cambios técnicos, normativos u organizativos.

CONTENIDOS

Unidad 1. Supervisión y monitorización

1. El dilema: más autonomía ≠ menos control
2. 4 niveles de supervisión; métricas (rendimiento, calidad, recursos, comportamiento)
3. Sistema de alertas; dashboard; logs y trazabilidad

Unidad 2. Guardrails: los 4 tipos

1. Filosofía: no confiar, verificar; diseñar para el peor caso
2. 4 tipos de guardrails: entrada (validar inputs, detectar manipulación), proceso (límites de pasos, timeouts, checkpoints), salida (validar output, confirmar irreversibles), recursos (tokens, rate limits, quotas)
3. Lista de comprobación pre-acción

Unidad 3. Seguridad de datos y contención

1. Protección de datos personales
2. Mitigación de prompt injection
3. Contención y respuesta ante fallos

Unidad 4. Escalado a humanos

1. Función del humano en el loop
2. Patrones de escalado operativo
3. Protocolos de intervención humana

Unidad 5. ACPC y auditoría

1. Fundamentos del marco ACPC
2. Capas de control aplicables
3. Evidencias y registros auditables
4. Requisitos de revisión continua

Unidad Práctica. Diseñar gobernanza de sistema agéntico