Hospital Real
Plaza Falla, 8 | 11003 Cádiz
Tel. 956 015 728
http://www.uca.es

GUÍA DOCENTE
ACTIVIDAD FORMATIVA PRESENCIAL

TÍTULO DE LA ACTIVIDAD		
RIESGOS DE LA IA GENERATIVA Y SEGURIDAD EN LA CREACIÓN DE AGENTES		
Duración (en horas)	**Nº Plazas**	**Campus/Lugar**
3	70	Salón de Grados Escuela Superior de Ingeniería Puerto Real
COORDINA:		Vicerrectorado de Transformación para la Universidad Digital
Correo electrónico		transformacion.digital@uca.es
Universidad/empresa/entidad		Universidad de Cádiz
FORMADOR/A		Miguel Ángel Hernández Ruiz
Correo electrónico		ingeniero.miguelruiz@gmail.com
Universidad/empresa/entidad		WHITETIE

Perfil de las personas destinatarias y requisitos a cumplir (según casos)

1. Equipos Directivos de Centros y Departamentos de la Universidad de Cádiz
2. PDI de la Universidad de Cádiz

PLANIFICACIÓN

Objetivos

- Comprender los principales riesgos del uso de la IA generativa en el entorno profesional.
- Identificar los ataques más habituales contra sistemas de IA generativa y agentes.
- Reconocer y mitigar los riesgos de seguridad al crear y compartir agentes en Copilot Studio.
- Aplicar buenas prácticas de gobernanza, cumplimiento y supervisión desde la dirección.
-

Contenidos

Módulo 0 Introducción y contexto

- Panorama de la IA generativa en la empresa: oportunidades y nueva superficie de riesgo.
- Por qué la seguridad de la IA es una responsabilidad que recae en la dirección.
- Objetivos y estructura de la sesión.

Módulo 1 Riesgos del uso de la IA generativa en el trabajo

- Datos sensibles y confidencialidad: fuga de información a través de los prompts; diferencia entre versiones de consumo y empresariales; uso de los datos para entrenamiento.

Hospital Real
Plaza Falla, 8 | 11003 Cádiz
Tel. 956 015 728
http://www.uca.es

- Sesgos y equidad: discriminación en decisiones (selección de personal, atención a clientes, crédito) y su impacto reputacional y legal.
    - Alucinaciones y fiabilidad: información inventada presentada con total seguridad; la necesidad de verificación humana antes de decidir.
    - Dependencia excesiva: pérdida de criterio profesional, atrofia de competencias y riesgo operativo ante la indisponibilidad de la herramienta.
    - Propiedad intelectual y derechos de autor: titularidad, licencias y uso de los contenidos generados.
    - «Shadow AI»: uso de herramientas de IA no autorizadas, fuera del control y la visibilidad de la organización.

Módulo 2 Ataques más comunes contra la IA generativa

- Inyección de prompts (directa e indirecta / XPIA): manipulación de las instrucciones del sistema a través de contenido aparentemente inofensivo.
- Fuga y exfiltración de información: cómo un atacante extrae datos sin acceso directo. Caso de referencia: «EchoLeak» en Microsoft 365 Copilot (CVE-2025-32711).
- Toma de control de agentes (navegador y autónomos): secuestro de objetivos («goal hijacking») y ejecución de acciones no deseadas en nombre del usuario.
- Envenenamiento de datos y memoria: corrupción del contexto o de las fuentes para condicionar las respuestas futuras del sistema.
- «Confused deputy» y abuso de permisos: el agente actúa con los privilegios de su creador en beneficio del atacante.
- Ingeniería social potenciada por IA: phishing, deepfakes y suplantación dirigidos específicamente a la dirección.

Módulo 3 Seguridad práctica en Copilot Studio: crear y compartir agentes

PUNTO DE PARTIDA

- Cómo la creación «no-code» amplía la superficie de ataque: cualquier persona puede crear agentes
- conectados a sistemas y datos corporativos.
- CASOS PRÁCTICOS Y DEMOSTRACIONES
- Inyección de prompts vía formularios / SharePoint: que desencadena la exfiltración de datos
- (investigación de Capsule Security; CVE-2026-21520).
- «Confused deputy»: los agentes se ejecutan con las credenciales de su creador y pueden exponer sus datos a cualquier usuario.
- Sobreexposición al compartir o publicar: agentes publicados en Teams, en la web o incluso accesibles de forma anónima sin autenticación.
- Backdoor mediante «Connected Agents»: exposición del conocimiento y las herramientas de un agente al resto de agentes del entorno (investigación de Zenity Labs).
- Phishing de OAuth desde el dominio legítimo («CoPhish»): abuso de la confianza en la propia plataforma de Microsoft.
- Conectores con exceso de permisos y brechas de DLP: flujos de datos que eluden las políticas de la organización.

BUENAS PRÁCTICAS DE CONFIGURACIÓN SEGURA

- Principio de mínimo privilegio en conectores y acceso a datos.
- Autenticación obligatoria y control estricto del alcance de la publicación.

Hospital Real
Plaza Falla, 8 | 11003 Cádiz
Tel. 956 015 728
http://www.uca.es

- Aplicación de políticas de prevención de pérdida de datos (DLP).
- Monitorización y auditoría de agentes (Microsoft Purview / Defender)

Módulo 4 Gobernanza, cumplimiento y responsabilidad directiva

- Política de uso aceptable y clasificación de datos: qué se puede y qué no se puede introducir en una
- herramienta de IA.
- Ciclo de vida del agente: aprobación, pruebas, publicación, revisión periódica y retirada.
- Marco normativo: RGPD, Reglamento Europeo de IA (EU AI Act) y deberes de confidencialidad y secreto.
- Rendición de cuentas y supervisión humana: el «human in the loop» y la responsabilidad última de las decisiones.
- Respuesta ante incidentes: qué hacer, y a quién avisar, cuando algo falla.

Módulo 5 Cierre, checklist y plan de acción

- Lista de verificación accionable para la dirección.
- Próximos pasos y recomendaciones.
- Preguntas y respuestas.

Metodología

Exposición dinámica apoyada en demostraciones en vivo y casos reales recientes, con espacio para preguntas a lo largo de toda la sesión. El enfoque prioriza la toma de decisiones y la gestión del riesgo desde una perspectiva directiva, más que el detalle técnico de implementación.

Tipo de certificación

X	**Asistencia**

Tutorías: sistemas de comunicación y atención al alumnado

La comunicación y atención se realizará a través de correo electrónico.

Calendario de sesiones

Sesiones	Campus/Lugar	Día	Horario
Sesión 1	Puerto Real Salón de Grados de la Escuela Superior de Ingeniería	05/10/2026	9:30 – 12:30