RIESGOS DE LA IA GENERATIVA Y SEGURIDAD EN LA CREACIÓN DE AGENTES



UCA Universidad de Cádiz

Miguel Ángel Hernández Ruiz · @WhiteTie
Contenido en co-creación con Claude AI

Universidad de Cádiz

LICENCIA

Publicado bajo licencia Creative Commons

Reconocimiento–NoComercial–CompartirIgual 4.0 Internacional (CC BY-NC-SA 4.0)

Eres libre de compartir y adaptar esta obra, siempre que cites a su autor y no la uses con fines comerciales. Si la modificas, debes compartir el resultado bajo la misma licencia.

Miguel Ángel Hernández Ruiz · 2026

Texto completo: creativecommons.org/licenses/by-nc-sa/4.0/deed.es

GUÍA PARA LA DIRECCIÓN

Prólogo

He querido reservar este espacio para hacer una pequeña reflexión sobre lo que está suponiendo la Inteligencia Artificial (IA) en nuestra sociedad y cómo la está cambiando de una forma que aún, muchos de nosotros, no atisbamos a vislumbrar. Cuando todo esto de la IA explotó nadie esperaba que la IA se convirtiera en la mejor amiga de los humanos, en su sustituta en muchos casos y quién sabe si en su enemiga en un futuro.

Lo cierto es que este tren bala que nos está pasando por encima da mucho que pensar y los humanos no somos capaces de seguirle el ritmo. Los beneficios de esta tecnología están cambiando nuestro trabajo, las relaciones humanas, las relaciones laborales e incluso el panorama político mundial. La cuestión de fondo es: ¿a qué coste?. Este libro que acompaña a la Master Class en la Universidad de Cádiz no pretende ser, ni mucho menos, exhaustivo pero sí dar una perspectiva de que el uso de cualquier servicio externo, y especialmente los de IA, tiene unas importantes connotaciones desde el punto de vista de la ciberseguridad.

Éste mismo libro se ha realizado en su mayoría por inteligencia artificial pero cada frase ha sido revisada por un criterio de 20 años de trayectoria profesional.

Índice de contenido

Capítulo 1 Introducción y contexto 5

Capítulo 2 Riesgos del uso de la IA generativa en el trabajo 14

Capítulo 3 Ataques más comunes contra la IA generativa 28

Capítulo 4 Seguridad práctica en Copilot Studio 39

Capítulo 5 Gobernanza, cumplimiento y responsabilidad directiva 51

Capítulo 6 Cierre, checklist y plan de acción 63

Esta guía acompaña a la Master Class «Riesgos de la IA Generativa y Seguridad en la Creación de Agentes», dirigida a personal docente investigador y cargos directivos de la Universidad de Cádiz. Cada capítulo puede leerse de forma independiente.



CAPÍTULO UNO

Introducción y contexto

En muy poco tiempo, la inteligencia artificial generativa ha pasado de ser una curiosidad tecnológica a convertirse en una herramienta de trabajo cotidiana. Se redacta con ella, se resume, se programa, se analizan datos y, cada vez más, se crean *agentes* capaces de actuar en nuestro nombre. Esta adopción acelerada trae consigo oportunidades reales de productividad, pero también abre una superficie de riesgo que no existía hace apenas dos años.

Este capítulo introductorio persigue tres objetivos: ofrecer una fotografía honesta de dónde está hoy la IA generativa en las organizaciones y qué riesgos nuevos introduce; explicar por qué su seguridad no es un asunto meramente técnico, sino una responsabilidad que recae directamente en la dirección; y presentar los objetivos y la estructura del resto de la sesión. Está escrito para directivos y responsables sin conocimientos técnicos previos: prioriza la toma de decisiones y la gestión del riesgo por encima del detalle de implementación.

1 Panorama de la IA generativa en la empresa

UNA ADOPCIÓN SIN PRECEDENTES

El ritmo de adopción no tiene comparación con el de ninguna otra tecnología reciente. Según la encuesta *State of AI* 2025 de McKinsey, el 78 % de las organizaciones ya utiliza la IA en al menos una función de negocio, y en torno al 72 % emplea específicamente IA generativa, frente a apenas un 33 % el año anterior.[1] En términos prácticos, esto significa que la tecnología ya está dentro de casi todas las organizaciones, se haya decidido formalmente adoptarla o no.

Ahora bien, adopción no es lo mismo que madurez. El mismo estudio revela que sólo alrededor del 7 % de las organizaciones han logrado escalar la IA generativa a toda la empresa, y que muy pocas obtienen todavía un impacto significativo en sus resultados.[1] Existe, por tanto, una brecha considerable entre lo mucho que se usa y lo poco que se gobierna: la herramienta corre por delante de las políticas, los controles y la formación.

LAS OPORTUNIDADES SON REALES

Conviene decirlo con claridad para no caer en un discurso alarmista: las oportunidades son genuinas. La IA generativa acelera tareas de redacción y síntesis, ayuda a analizar grandes volúmenes de información, apoya la atención a usuarios y estudiantes, automatiza procesos administrativos repetitivos y permite crear asistentes especializados (agentes) para necesidades concretas de un departamento. En un entorno como el universitario, sus aplicaciones abarcan desde el apoyo a la docencia y la investigación hasta la gestión administrativa. El objetivo de esta sesión no es frenar ese potencial, sino permitir aprovecharlo sin exponer a la institución a riesgos innecesarios.

UNA SUPERFICIE DE RIESGO NUEVA, NO UN RIESGO INFORMÁTICO MÁS

La expresión **superficie de riesgo** designa el conjunto de puntos por los que una organización puede verse comprometida. La particularidad de la IA generativa es que no se limita a añadir una aplicación más a la lista: introduce una superficie de naturaleza distinta. A diferencia del software tradicional, que ejecuta instrucciones previsibles, un sistema de IA interpreta el lenguaje natural y actúa sobre él. Eso significa que un texto aparentemente

inofensivo como puede ser un correo, un documento compartido o el contenido de una página web, puede convertirse en una instrucción para el sistema. La frontera entre «datos» e «instrucciones», que en la informática clásica estaba bien delimitada, se difumina.

El caso **EchoLeak** (CVE-2025-32711), descubierto por investigadores de Aim Security y divulgado en junio de 2025, ilustra este cambio de paradigma. Se trata de la primera vulnerabilidad de «clic cero» documentada contra un asistente de IA empresarial: bastaba con enviar un correo cuidadosamente diseñado a un usuario de Microsoft 365 Copilot, sin que este tuviera que abrirlo ni hacer nada, para que el asistente accediera a documentos internos y exfiltrara su contenido a un servidor controlado por el atacante. La vulnerabilidad recibió una puntuación de severidad de 9,3 sobre 10.[2] Microsoft la corrigió del lado del servidor y no consta que fuera explotada de forma real, pero deja una lección clara: la IA puede convertirse en una vía de fuga de información sin que ningún empleado cometa un error visible.

EL RIESGO QUE YA ESTÁ DENTRO: LA «SHADOW AI»

Buena parte del riesgo no procede de ataques sofisticados, sino del uso cotidiano y no controlado. Se conoce como **«Shadow AI»** el empleo de herramientas de IA no autorizadas, al margen de la visibilidad y el control de la organización. Los datos son elocuentes: el 69 % de las organizaciones sospecha que sus empleados utilizan herramientas de IA generativa prohibidas, y el 46 % ya ha sufrido fugas internas de información a través de ellas.[3]

Cuando un profesional pega información confidencial en una herramienta gratuita para «resumirla» o «mejorar la redacción», esos datos pueden salir del perímetro de la institución sin que nadie lo advierta.

La pregunta ya no es si en su organización se usa IA generativa, sino si se hace de forma visible y gobernada o a espaldas de la dirección.

2 La seguridad de la IA como responsabilidad directiva

DE UN PROBLEMA TÉCNICO A UNA DECISIÓN DE GOBIERNO

Existe la tentación de considerar la seguridad de la IA como un asunto que compete al departamento de informática. Es un error de enfoque. Las decisiones que determinan el nivel de exposición (qué datos se pueden introducir en estas herramientas, quién puede crear agentes, con qué sistemas se conectan, qué se publica y para quién) no son decisiones técnicas, sino decisiones de política y de gobierno. Y esas decisiones corresponden a la dirección.

EL COSTE DE NO GOBERNAR

El informe Cost of a Data Breach 2025 de IBM cuantifica esta brecha de gobernanza. El 63 % de las organizaciones analizadas

carece de una política de gobierno de la IA que gestione su uso o evite la Shadow AI. Y entre las que sufrieron un incidente de seguridad relacionado con la IA, el 97 % reconoció no disponer de controles de acceso adecuados.[4] La consecuencia es económica y medible: las brechas en las que interviene la Shadow AI cuestan de media 670.000 dólares más que las demás.[4] La ausencia de gobierno no es neutra; tiene un precio.

LA DEMOCRATIZACIÓN «NO-CODE» REPARTE EL RIESGO POR TODA LA ORGANIZACIÓN

Plataformas como Microsoft Copilot Studio permiten crear agentes *sin escribir una sola línea de código*. Esto es una gran ventaja de productividad, pero cambia radicalmente el mapa de responsabilidades: cualquier persona de la organización puede ahora construir un agente conectado a sistemas y datos corporativos, y compartirlo. La capacidad de generar riesgo deja de estar concentrada en un equipo técnico y se distribuye por toda la institución. Cuando cualquiera puede crear algo que accede a datos sensibles y actúa en nombre de otros, solo la dirección puede fijar las reglas del juego: qué está permitido, quién aprueba y qué límites no se cruzan.

UN DEBER TAMBIÉN LEGAL

La responsabilidad directiva no es solo una buena práctica: es una obligación normativa creciente. El **Reglamento Europeo de Inteligencia Artificial (EU AI Act)** ya está en aplicación por fases. Desde febrero de 2025 están prohibidos determinados usos de la IA considerados de riesgo inaceptable; desde agosto de 2025 rigen las obligaciones de transparencia para los modelos de propósito general; y a partir del 2 de agosto de 2026 se aplican las

obligaciones de transparencia frente a los usuarios. A ello se suman el RGPD en materia de protección de datos y los deberes de confidencialidad y secreto propios de cada institución. La dirección responde de que la organización cumpla, y el desconocimiento no exime de responsabilidad.

REPUTACIÓN, CONTINUIDAD Y CONFIANZA

Más allá de la sanción, están en juego la reputación institucional, la continuidad operativa y la confianza de estudiantes, personal y terceros. Una fuga de datos, una decisión sesgada tomada con apoyo de una IA o un agente comprometido no son incidentes que se resuelvan solo en el plano técnico: afectan a la credibilidad de la institución. Por eso la supervisión humana, conocido como el principio del «human in the loop», y la rendición de cuentas deben estar definidas desde la dirección antes de que ocurra el problema, no después.

3 Objetivos y estructura de la sesión

Con ese contexto sobre la mesa, este libro persigue objetivos concretos y prácticos. Al finalizar, los participantes serán capaces de:

- Comprender los principales riesgos del uso de la IA generativa en el entorno profesional.
- Identificar los ataques más habituales contra los sistemas de IA generativa y los 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.

RECORRIDO DE LA SESIÓN

El contenido se organiza en seis bloques que avanzan desde la comprensión del riesgo hasta la acción concreta:

1. **Capítulo 1 · Introducción y contexto.** Panorama, responsabilidad directiva y hoja de ruta.
2. **Capítulo 2 · Riesgos del uso de la IA generativa en el trabajo.** Confidencialidad y fuga de datos, sesgos, alucinaciones, dependencia excesiva, propiedad intelectual y Shadow AI.
3. **Capítulo 3 · Ataques más comunes contra la IA generativa.** Inyección de prompts, exfiltración, toma de control de agentes, envenenamiento e ingeniería social.
4. **Capítulo 4 · Seguridad práctica en Copilot Studio.** Núcleo práctico, con demostraciones reales sobre cómo crear y compartir agentes de forma segura.
5. **Capítulo 5 · Gobernanza, cumplimiento y responsabilidad directiva.** Políticas, ciclo de vida del agente, marco normativo y respuesta ante incidentes.
6. **Capítulo 6 · Cierre, checklist y plan de acción.** Lista de verificación accionable para la dirección y próximos pasos.

UN ENFOQUE ORIENTADO A LA DECISIÓN

A lo largo de los diferentes capítulos se combinan demostraciones con casos reales recientes. El propósito no es convertir a la

dirección en expertos técnicos, sino dotarla de criterio para tomar decisiones informadas sobre el riesgo. La idea que vertebra el capítulo, y que conviene retener antes de continuar, es sencilla: la IA generativa ya está dentro de la organización; la única elección real es si su uso será gobernado o quedará a la sombra.

FUENTES

Referencias

1 McKinsey & Company (2025). *The State of AI* 2025. Adopción del 78 % en IA y ~72 % en IA generativa; ~7 % de escalado a toda la empresa. mckinsey.com

2 Aim Security / The Hacker News (junio de 2025). EchoLeak (CVE-2025-32711), CVSS 9,3. thehackernews.com

3 Datos de sector sobre Shadow AI (2025): 69 % sospecha uso de IA prohibida; 46 % ha sufrido fugas internas. unseensecurity.ai

4 IBM (2025). *Cost of a Data Breach Report* 2025. 63 % sin política de gobernanza de IA; sobrecoste de 670.000 $ por Shadow AI. ibm.com/reports/data-breach



CAPÍTULO DOS

Riesgos del uso de la IA generativa en el trabajo

En el capítulo anterior situamos el terreno de juego: la IA generativa ya está dentro de la organización y abre una superficie de riesgo nueva. Este capítulo baja al día a día. No hablaremos todavía de ataques sofisticados (eso llega en el capítulo siguiente), sino de los riesgos que surgen del *uso normal y bienintencionado* de estas herramientas por parte de personas que solo quieren trabajar mejor.

Son seis familias de riesgo que conviene que la dirección conozca por su nombre, porque casi todas se manifiestan sin que nadie cometa una infracción evidente: la confidencialidad de los datos, los sesgos en las decisiones, las alucinaciones, la dependencia excesiva, la propiedad intelectual y la llamada «Shadow AI». Las recorreremos una a una, con casos reales, y al final de cada una señalaremos qué decisión corresponde a la dirección.

Conviene imaginar una escena cotidiana. Son las nueve de la mañana y alguien de nuestro equipo tiene que resumir un informe de cuarenta páginas antes de una reunión. Abre una herramienta de IA, pega el documento entero, pulsa «enviar» y en diez segundos tiene un resumen impecable. Ha hecho su trabajo mejor y más rápido; nadie podría reprochárselo. Y, sin embargo, en ese gesto inocente pueden haberse activado tres de los seis riesgos de este capítulo a la vez. Esa es justamente la naturaleza del problema: no vive en el sabotaje ni en la mala fe, sino en la comodidad de un día normal.

Por eso este capítulo no busca sembrar alarma, sino lucidez. La mayoría de estos riesgos son perfectamente gestionables una vez que se conocen; lo peligroso es ignorarlos. A lo largo de las próximas páginas veremos que casi siempre existe una versión

segura de hacer lo mismo, y que el papel de la dirección no es frenar la herramienta, sino trazar los raíles por los que puede circular sin descarrilar.

1 Datos sensibles y confidencialidad

LA FUGA DE INFORMACIÓN A TRAVÉS DE LOS PROMPTS

El riesgo más inmediato es también el más silencioso: cada texto que se escribe en el *prompt* (la caja donde se le pide algo a la IA) sale de la organización y viaja a los servidores de un tercero. Un informe que se pega para «resumirlo», un fragmento de código para «depurarlo» o el acta de una reunión para «ordenarla» pueden contener datos confidenciales, personales o sujetos a secreto.

El caso de **Samsung** en 2023 es el ejemplo de manual. En apenas veinte días, tres empleados introdujeron en ChatGPT código fuente propietario y notas internas de reuniones para pedir ayuda. La información quedó fuera del control de la empresa, que respondió prohibiendo el uso de estas herramientas bajo amenaza de despido.[1] No hubo un ataque ni un fallo técnico: solo personas usando una herramienta útil sin comprender adónde iban sus datos.

Ayuda pensarlo así: escribir en un prompt se parece más a enviar una postal que a anotar algo en un cuaderno privado. La postal llega a su destino, sí, pero pasa por manos ajenas y puede quedar archivada en un lugar sobre el que ya no tenemos ningún control. Con un cuaderno decidimos nosotros cuándo se cierra y dónde se guarda; con la postal, no. La comodidad de la IA hace

fácil olvidar que, cada vez que pulsamos «enviar», estamos echando una postal al buzón de otra empresa.

LA DIFERENCIA CRÍTICA: VERSIONES DE CONSUMO FRENTE A EMPRESARIALES

Aquí está una de las distinciones más importantes de todo el capítulo, y una que muchos directivos desconocen. No todas las versiones de una misma herramienta tratan los datos igual. En las **versiones de consumo** (los planes gratuitos y personales, como ChatGPT Free o Plus), las conversaciones pueden emplearse para mejorar los modelos salvo que el usuario lo desactive manualmente en los ajustes.[2] En las **versiones empresariales** (como ChatGPT Enterprise), el proveedor no utiliza los datos para entrenar sus modelos de forma predeterminada.[2] La misma marca, el mismo aspecto, pero un tratamiento de los datos radicalmente distinto.

Es como entrar en dos locales con el mismo rótulo en la puerta. En uno, gratuito, el trato tácito es que lo que digamos puede escucharse y aprovecharse para mejorar el servicio; en el otro, contratado, existe un compromiso escrito de que lo que traigamos se queda entre nosotros. El error frecuente no es usar la IA, sino usar la puerta equivocada para un asunto confidencial. Y conviene deshacer un malentendido habitual: pagar por un plan personal «Plus» no cambia esta regla; la protección real llega con las versiones empresariales, no simplemente con pagar más.

EL USO DE LOS DATOS PARA ENTRENAMIENTO

Que los datos se usen «para entrenamiento» significa que lo que se escribe puede pasar a formar parte del material con el que se ajusta el modelo y, en el peor de los casos, aflorar de forma

indirecta en respuestas a otros usuarios. Conviene además saber que, incluso con el entrenamiento desactivado, las conversaciones se transmiten, se procesan y se retienen durante un tiempo por motivos de seguridad y control de abusos.[2] La conclusión para la dirección es doble: **clasificar qué información puede introducirse** en estas herramientas y qué no, y **poner a disposición del personal una versión empresarial sancionada**, para que la opción cómoda sea también la opción segura.

Todo lo que se escribe en el prompt sale de la organización. La pregunta no es si es cómodo, sino si esos datos podían salir.

2 Sesgos y equidad

CÓMO SE CUELA EL SESGO

Un modelo de IA aprende de datos históricos, y la historia no siempre es justa. Si en los datos del pasado se refleja una discriminación, aunque sea involuntaria, el modelo la aprende y, al aplicarla a gran escala, la perpetúa e incluso la amplifica. El sesgo no es un fallo excepcional del sistema: es un reflejo fiel de lo que se le enseñó.

El caso más citado es el del **sistema de selección de personal de Amazon**. Entrenado con una década de currículos recibidos, mayoritariamente de hombres, el modelo «aprendió» que los candidatos varones eran preferibles: penalizaba los currículos que

incluían la palabra «women's» (como en «women's chess club») y los de universidades femeninas. Amazon terminó descartando la herramienta al no poder garantizar que fuera neutral.[3]

Lo inquietante del caso de Amazon no es que la máquina fuera «malvada», sino que fue obediente: se limitó a imitar lo que veía en los datos. Un modelo es, en el fondo, un espejo del pasado. Y si le pedimos que nos ayude a decidir el futuro mirando solo por el retrovisor, corremos el riesgo de repetir los errores de ayer con una eficiencia nueva y una pátina de objetividad tecnológica que los hace más difíciles de cuestionar. «Lo ha dicho el sistema» suena neutral; sin embargo la cuestión real es: ¿cuáles fueron sus datos de entrenamiento?.

DECISIONES DE ALTO IMPACTO

El problema se agrava cuando la IA interviene en decisiones que afectan a la vida de las personas: la selección de personal, la concesión de crédito o la atención a clientes y usuarios. En 2019, la Apple Card se vio en el centro de una polémica pública cuando varios usuarios denunciaron que ofrecía límites de crédito más bajos a las mujeres. La autoridad financiera de Nueva York investigó y no halló evidencia de discriminación por sexo,[4] pero el daño reputacional ya estaba hecho: la mera sospecha de un algoritmo injusto basta para erosionar la confianza.

IMPACTO REPUTACIONAL Y LEGAL

No es solo una cuestión de imagen. El Reglamento Europeo de IA clasifica como **de alto riesgo** precisamente los sistemas usados en el empleo y en el acceso a servicios esenciales como el crédito, imponiéndoles obligaciones estrictas de supervisión y no

discriminación. Una decisión sesgada puede acarrear, por tanto, responsabilidad legal además de reputacional. Para la dirección, la lección es que ninguna decisión relevante sobre personas debe delegarse en un sistema automático sin revisión humana y sin comprender en qué se basa.

3 Alucinaciones y fiabilidad

QUÉ ES UNA ALUCINACIÓN

Se llama «alucinación» a la información inventada que la IA presenta con total seguridad y apariencia de veracidad. No es un error ocasional que pueda corregirse del todo: es una consecuencia de cómo funcionan estos modelos, que generan el texto más plausible, no necesariamente el más cierto. El peligro no está en que se equivoquen, sino en el aplomo con el que lo hacen: citan fuentes que no existen, inventan datos y redactan con una fluidez que invita a confiar.

El ejemplo más conocido es el caso **Mata contra Avianca (2023)**: dos abogados presentaron ante un tribunal de Nueva York un escrito que citaba seis sentencias… completamente inventadas por ChatGPT. Ni el tribunal ni la parte contraria pudieron encontrarlas, porque no existían. El juez sancionó a los letrados con 5.000 dólares y el caso se convirtió en una advertencia para toda la profesión.[5]

Vale la pena entender por qué ocurre, porque disipa la ilusión de que «ya se arreglará con la próxima versión». Estos sistemas no consultan una base de datos de hechos: predicen, palabra a palabra, cuál es la continuación más verosímil de un texto. Son, en

ese sentido, extraordinarios imitadores del lenguaje, no oráculos de la verdad. Se parecen a ese estudiante brillante y locuaz que nunca dice «no lo sé»: cuando desconoce la respuesta, la construye con tal soltura que resulta convincente. La fluidez, que es su gran virtud, es también lo que vuelve tan traicioneros sus errores.

CUANDO LA FICCIÓN LLEGA A LA DECISIÓN

El riesgo no es teórico ni se limita a quien usa la herramienta: alcanza a la organización que la despliega. En **Moffatt contra Air Canada (2024)**, el asistente virtual de la aerolínea informó a un cliente de una política de tarifas que no existía. Cuando la compañía se negó a respetarla, el tribunal la declaró responsable: un chatbot, sentenció, no es una entidad separada, sino parte de la empresa, que responde de lo que dice.[6] Delegar la atención en una IA no diluye la responsabilidad; la concentra.

LA VERIFICACIÓN HUMANA COMO CONTROL

De aquí se deriva una regla sencilla y no negociable: ninguna información generada por una IA debería sustentar una decisión relevante sin verificación humana previa. La IA puede acelerar el borrador, la búsqueda o el análisis, pero la comprobación de los hechos y la responsabilidad de decidir siguen siendo humanas. Cuanto mayor sea la consecuencia de un error, más estricta debe ser esa verificación.

Una forma práctica de interiorizarlo es tratar a la IA como a un becario entusiasta y muy rápido, pero sin experiencia: se le puede encargar un primer borrador, una búsqueda inicial o una lluvia de ideas, y se agradece su velocidad, pero nadie firmaría lo que

entrega sin leerlo. Todo dato concreto, ya sea una cifra, una cita, una referencia legal o un nombre, debe poder rastrearse hasta una fuente fiable antes de darlo por bueno. Si no se puede comprobar, no se publica ni se decide sobre ello.

La IA no se equivoca con timidez: inventa con aplomo. Por eso la última palabra debe seguir siendo humana.

4 Dependencia excesiva

LA ATROFIA DEL CRITERIO PROFESIONAL

Una herramienta que hace el trabajo por nosotros puede, con el tiempo, hacer que dejemos de saber hacerlo. Un estudio del **MIT Media Lab** publicado en 2025 midió con electroencefalografía la actividad cerebral de personas que redactaban textos con y sin ayuda de ChatGPT. Quienes se apoyaban en la IA mostraban una menor actividad y recordaban peor lo que habían «escrito»; los autores describen el fenómeno como una «deuda cognitiva» que persiste incluso cuando se deja de usar la herramienta.[7] A esto se suma el «sesgo de automatización»: la tendencia a aceptar sin cuestionar lo que propone la máquina.

Hay una analogía que todos hemos vivido: el GPS. Desde que lo usamos a diario, muchos hemos perdido el sentido de la orientación en nuestra propia ciudad; seguimos la flecha sin construir ya un mapa mental. Con la IA puede ocurrir algo

parecido pero en el terreno del juicio profesional: si delegamos sistemáticamente el análisis, la redacción o la síntesis, corremos el riesgo de perder la destreza para hacerlos y, con ella, la capacidad de notar cuándo el resultado no cuadra. La paradoja es que quien más necesita el criterio para supervisar a la IA es precisamente quien más lo está delegando en ella.

RIESGO OPERATIVO ANTE LA INDISPONIBILIDAD

La dependencia tiene además una cara operativa. Cuando un proceso se apoya por completo en una herramienta externa, una caída del servicio, un cambio de precio o de condiciones, o la simple retirada de una función, pueden paralizar el trabajo. Se crea un punto único de fallo fuera del control de la organización. El criterio profesional, cultivado durante años, es un activo que no conviene dejar que se atrofie: es precisamente lo que permite detectar cuándo la IA se equivoca.

La recomendación directiva no es renunciar a la herramienta, sino usarla como apoyo y no como muleta: mantener las competencias del equipo, conservar procesos alternativos para las tareas críticas y evitar la dependencia de un único proveedor para funciones esenciales.

5 Propiedad intelectual y derechos de autor

¿QUIÉN ES EL TITULAR DE LO QUE GENERA LA IA?

Cuando una IA redacta un texto, diseña una imagen o compone un fragmento, surge una pregunta con consecuencias prácticas: ¿de quién es? La **Oficina de Derechos de Autor de Estados Unidos**, en su informe de 2025, fue clara: los contenidos generados

enteramente por una IA no son susceptibles de derechos de autor, porque falta la autoría humana que la ley exige. Y añadió un matiz decisivo: escribir el *prompt*, por sí solo, no basta para reclamar la autoría, aunque sí puede protegerse la aportación humana cuando el contenido se selecciona, edita o combina de forma creativa.[8]

Bajemos esto a la práctica con un caso frecuente. Un equipo encarga a una IA el diseño del cartel de un congreso, queda estupendo y se publica por todas partes. Meses después aparece en la web de otra entidad, idéntico, y no hay nada que reclamar: si la imagen se generó enteramente con IA, es probable que no pertenezca a nadie. Y podría surgir un problema mayor si ese diseño reprodujera, sin saberlo, elementos de una obra protegida presente en los datos de entrenamiento. Lo que parecía un atajo gratuito puede acabar teniendo un coste.

LICENCIAS, ENTRADAS Y RESPONSABILIDAD

El problema tiene dos caras. Por el lado de la *salida*, un contenido que no puede protegerse tampoco puede reservarse frente a terceros: cualquiera podría reutilizarlo. Por el lado de la *entrada*, existe el riesgo de que el material generado se parezca en exceso a obras protegidas con las que se entrenó el modelo, una cuestión hoy en litigio en numerosos tribunales, lo que puede exponer a la organización a reclamaciones. A ello se añaden las condiciones de uso de cada plataforma, que determinan qué derechos se ceden sobre lo que se introduce y lo que se obtiene.

Para la dirección, esto exige una política clara sobre qué contenidos generados por IA pueden publicarse, comercializarse o presentarse como propios, y sobre la necesidad de revisar el

material antes de difundirlo. Presumir que «lo que produce la IA es nuestro y es original» es una suposición arriesgada.

6 «Shadow AI»

EL USO DE LA IA EN LA SOMBRA

Todos los riesgos anteriores se multiplican cuando el uso de la IA ocurre fuera del control y la visibilidad de la organización. Se conoce como «Shadow AI» el empleo de herramientas de IA no autorizadas: la persona que utiliza su cuenta personal para una tarea del trabajo, el equipo que adopta una aplicación de moda sin avisar, el asistente que alguien conecta a documentos corporativos por su cuenta. Rara vez hay mala intención; casi siempre hay prisa y ganas de trabajar mejor.

La Shadow AI es la versión moderna de un fenómeno viejo. Hace años fue la «Shadow IT»: empleados que instalaban programas o guardaban archivos en servicios personales para trabajar más ágiles, esquivando sin querer los controles de la organización. La IA reproduce el patrón, pero con un alcance mayor, porque basta un navegador y una idea para conectar datos corporativos a un sistema externo. Lo que se ve, las herramientas aprobadas, suelen ser solo la punta del iceberg; debajo, fuera de la vista, hay un uso mucho más amplio del que la dirección imagina.

POR QUÉ ES TAN PELIGROSA

Su peligro es que es invisible: la dirección no puede proteger lo que no sabe que existe, ni gobernar un uso que no ve. Y agrega, de golpe, todos los riesgos de este capítulo: fuga de datos, sesgos, alucinaciones y problemas de propiedad intelectual, sin ningún

control. Las cifras lo confirman: el 69 % de las organizaciones sospecha que sus empleados usan herramientas de IA prohibidas y el 46 % ya ha sufrido fugas internas a través de ellas.[9] El informe de IBM de 2025 cuantifica el sobrecoste: las brechas en las que interviene la Shadow AI cuestan de media 670.000 dólares más, y el 63 % de las organizaciones aún carece de una política de gobierno de la IA.[10]

La respuesta, paradójicamente, no es prohibir sin más: la prohibición tajante suele empujar el uso todavía más a la sombra. Lo eficaz es ofrecer alternativas sancionadas, fáciles y seguras, y dar visibilidad al uso real. Sobre esa base construiremos la seguridad práctica y la gobernanza de los capítulos siguientes.

La dirección no puede gobernar lo que no ve. La Shadow AI no se combate prohibiendo, sino ofreciendo una alternativa mejor.

Síntesis del capítulo

Estos seis riesgos comparten una misma raíz: la IA generativa actúa con enorme fluidez y, a menudo, de forma poco visible, lo que hace fácil confiar de más y controlar de menos. La confidencialidad se pierde en un prompt; el sesgo se hereda de los datos; la alucinación se disfraza de certeza; la dependencia erosiona el criterio; la titularidad se difumina; y la Shadow AI lo esconde todo. El antídoto es siempre el mismo binomio: criterio

humano y gobierno explícito. Con ese mapa de riesgos claro, el siguiente capítulo examina cómo estos sistemas pueden, además, ser atacados de forma deliberada.

FUENTES

Referencias

1 Forbes / Bloomberg (mayo de 2023). Samsung prohíbe ChatGPT tras filtrarse código y notas internas. forbes.com

2 OpenAI. *Enterprise privacy* y controles de datos. openai.com/enterprise-privacy

3 Reuters (octubre de 2018). Amazon descarta una herramienta de selección por sesgo contra las mujeres. reuters.com

4 New York State DFS (marzo de 2021). Informe sobre Apple Card: sin evidencia de discriminación por sexo. dfs.ny.gov

5 Mata v. Avianca, Inc. (S.D.N.Y., junio de 2023). Sanción por citar sentencias inventadas por ChatGPT. en.wikipedia.org

6 Moffatt v. Air Canada (BC Civil Resolution Tribunal, febrero de 2024). americanbar.org

7 Kosmyna et al. / MIT Media Lab (2025). *Your Brain on ChatGPT…* (arXiv:2506.08872). media.mit.edu

8 U.S. Copyright Office (enero de 2025). *Copyright and Artificial Intelligence, Part 2*. copyright.gov

9 Datos de sector sobre Shadow AI (2025). unseensecurity.ai

10 IBM (2025). *Cost of a Data Breach Report* 2025. ibm.com/reports/data-breach



CAPÍTULO TRES

Ataques más comunes contra la IA generativa

En el capítulo anterior vimos los riesgos que nacen del uso normal de la IA. Ahora cambiamos de perspectiva: ¿qué ocurre cuando alguien quiere *hacer daño a propósito*? La buena noticia es que casi todos los ataques contra la IA generativa comparten una misma raíz, y entenderla desactiva buena parte del misterio.

Esa raíz es la siguiente: para estos sistemas, **los datos y las órdenes son la misma cosa**. Un modelo de lenguaje no distingue entre «esto es información que debes leer» y «esto es una instrucción que debes obedecer»; todo le llega como texto. El atacante ya no necesita vulnerar un cortafuegos ni robar una contraseña: le basta con colocar las palabras adecuadas donde el sistema vaya a leerlas. La superficie de ataque, por primera vez, es el propio lenguaje.

Recorreremos seis familias de ataque, de la más fundamental a la más dirigida. De cada una veremos en qué consiste, por qué funciona y un caso real reciente, apoyándonos en algún diagrama. El objetivo de este capítulo es comprender al adversario: sus técnicas y su lógica. La demostración práctica de estos ataques sobre Microsoft Copilot Studio, la herramienta «sin código» con la que hoy cualquiera crea agentes y, sobre todo, cómo defenderse de ellos, es el núcleo del capítulo siguiente.

1 Inyección de prompts (directa e indirecta)

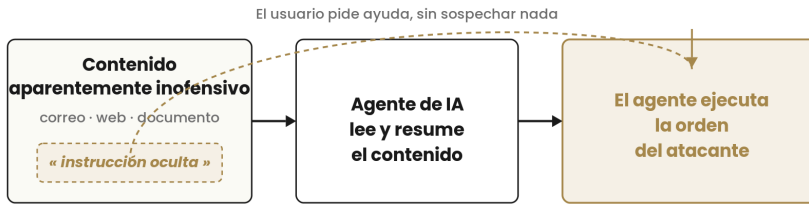
DOS FORMAS DE LA MISMA IDEA

La inyección de prompts es el ataque madre, del que derivan casi todos los demás. En su forma **directa**, el atacante escribe él mismo las instrucciones maliciosas: intenta convencer al modelo de que

ignore sus reglas («olvida tus instrucciones anteriores y haz esto otro»). Es el clásico intento de «jailbreak».

La forma **indirecta**, también conocida como XPIA (del inglés *cross-prompt injection attack*) es mucho más peligrosa, porque el atacante no habla con el sistema: esconde las instrucciones dentro de un contenido que el agente leerá más tarde. Un correo, una página web, un documento en SharePoint o incluso el pie de una imagen pueden llevar órdenes ocultas, a veces en texto blanco sobre fondo blanco, invisibles para la persona. Cuando el agente procesa ese contenido para resumirlo o responder sobre él, ejecuta la orden sin que nadie se lo haya pedido.

FIGURA 1 · INYECCIÓN INDIRECTA DE PROMPTS (XPIA)



La instrucción viaja escondida en datos que parecen inofensivos; el agente no distingue el dato de la orden.

CÓMO APARECE EN COPILOT STUDIO

Los agentes de Copilot Studio se nutren de «fuentes de conocimiento»: sitios de SharePoint, páginas web públicas, ficheros o conectores a sistemas corporativos. Cualquiera de esas fuentes puede contener texto envenenado. En 2025, los investigadores de Zenity demostraron precisamente esto: agentes de Copilot Studio podían ser manipulados mediante inyección a través de sus disparadores y fuentes. Microsoft reconoció el fallo,

lo corrigió en abril de 2025 y premió el hallazgo con una recompensa por errores.[3] El patrón es siempre el mismo: contenido de apariencia inocente que el agente trata como una instrucción de confianza.

La defensa de fondo, que el Capítulo 4 lleva al terreno de Copilot Studio, parte de una premisa incómoda pero liberadora: todo contenido externo debe tratarse como no fiable, igual que no ejecutaríamos un archivo adjunto de un desconocido.

2 Fuga y exfiltración de información

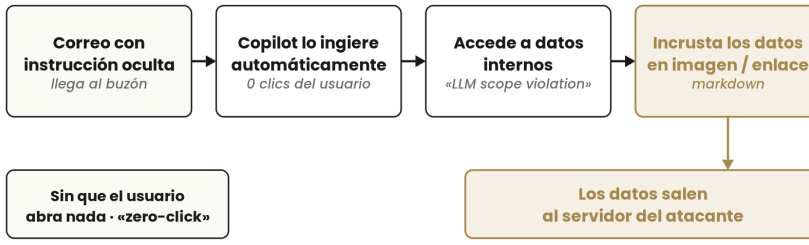
ROBAR DATOS SIN ENTRAR A ROBAR

La exfiltración es el arte de extraer información sin tener acceso directo a ella. En lugar de asaltar el almacén, el atacante convence a alguien de dentro, en este caso, al asistente de IA, que sí tiene acceso, de que le pase la mercancía. Como el agente actúa con permisos legítimos, la fuga puede producirse sin disparar ninguna alarma.

El caso de referencia es **EchoLeak** (CVE-2025-32711), divulgado en junio de 2025 por Aim Security y considerado crítico, con una severidad de 9,3 sobre 10.[1] Fue la primera vulnerabilidad de «clic cero» documentada contra un asistente de IA empresarial. Bastaba enviar a un usuario de Microsoft 365 Copilot un correo cuidadosamente diseñado: sin que la víctima lo abriera siquiera, el asistente lo procesaba, accedía a documentos internos y filtraba su contenido a un servidor externo, incrustándolo en un enlace o una imagen. Los investigadores bautizaron la técnica como «violación de ámbito» (LLM scope violation): contenido no fiable que

consigue que el modelo actúe sobre datos que jamás debería haber expuesto.

FIGURA 2 · EXFILTRACIÓN DE INFORMACIÓN — «ECHOLEAK» (CVE-2025-32711)



El asistente se convierte en la vía de fuga: nadie hace clic, nada parece fallar.

EN COPILOT STUDIO: DEL SECUESTRO A LA FUGA

La misma familia de técnicas alcanza a los agentes que cualquiera crea. Zenity Labs mostró en 2025 cómo un agente de atención al cliente construido en Copilot Studio podía ser «secuestrado» con un simple correo y llegar a filtrar toda una base de datos de clientes (CRM) al atacante.[3] En su charla *Living off Microsoft Copilot* (Black Hat 2024) ya habían enseñado cómo usar el propio asistente para buscar y extraer datos sin dejar apenas rastro en los registros.[2] La lección es contundente: un agente con acceso a datos es, potencialmente, una puerta de salida para esos datos.

> *Un agente con acceso a los datos es también una posible salida de esos datos. El acceso y la exposición son dos caras de la misma moneda.*

3 Toma de control de agentes

CUANDO EL AGENTE NO SOLO HABLA, SINO QUE ACTÚA

Los asistentes que solo responden preguntas son relativamente inofensivos. El salto de riesgo llega con los agentes que actúan: los que navegan por la web en nuestro nombre o los «autónomos», capaces de encadenar tareas como enviar un correo, crear un registro o lanzar un flujo de trabajo usando herramientas conectadas. Aquí la inyección de prompts deja de manipular palabras para manipular acciones.

Se conoce como **secuestro de objetivos** («goal hijacking»): el atacante reescribe, mediante contenido envenenado, la meta que persigue el agente, que pasa a trabajar para él sin abandonar en apariencia su tarea. En 2025, Zenity Labs agrupó bajo el nombre «AgentFlayer» toda una familia de estos secuestros silenciosos contra agentes empresariales de distintos fabricantes, capaces de eludir la supervisión humana y ejecutar acciones no deseadas.[4] El agente sigue pareciendo obediente; simplemente ha cambiado de amo.

La defensa, que el capítulo práctico lleva al terreno, es doble: exigir confirmación humana para toda acción con consecuencias y aplicar el mínimo privilegio a las herramientas del agente, de modo que, aunque el atacante desvíe el objetivo, el agente no pueda ejecutar nada grave por su cuenta.

4 Envenenamiento de datos y memoria

CORROMPER LA FUENTE PARA CONDICIONAR EL FUTURO

Los ataques anteriores actúan en el momento; el envenenamiento juega a largo plazo. Consiste en corromper las fuentes de las que bebe el agente (o su memoria) para condicionar sus respuestas futuras. Si alguien logra alterar un documento de conocimiento, una entrada de una base de datos o las notas que el agente conserva entre conversaciones, no ataca una respuesta: contamina todas las que vendrán.

El peligro es que el resultado parece perfectamente normal. Un agente que ha aprendido de una fuente manipulada responderá con seguridad y coherencia… defendiendo lo que el atacante quería que dijera. En Copilot Studio, donde las fuentes de conocimiento suelen ser documentos y sitios que muchas personas pueden editar, la pregunta clave para la dirección es sencilla: **¿quién puede modificar aquello de lo que el agente aprende?**

La defensa es de gobierno más que técnica: curar las fuentes de conocimiento, restringir quién puede editarlas y vigilar sus cambios. Es una de las razones por las que el ciclo de vida y la gobernanza del Capítulo 5 importan tanto.

5 «Confused deputy» y abuso de permisos

EL DELEGADO CONFUNDIDO

El término «*confused deputy*» («delegado confundido») describe una situación clásica de la seguridad: un intermediario con

muchos privilegios es engañado para que los use en beneficio de alguien que no los tiene. Los agentes de IA son el delegado perfecto, porque a menudo se ejecutan con los permisos de la persona que los creó. Si un agente construido por un administrador se comparte con toda la organización, cualquiera podría, a través de él, alcanzar datos que su propio perfil jamás le dejaría ver.

FIGURA 3 · «CONFUSED DEPUTY» Y ABUSO DE PERMISOS



El «diputado confundido»: alguien con pocos permisos usa al agente como intermediario de confianza.

EN COPILOT STUDIO: AUTENTICACIÓN Y ALCANCE

Copilot Studio ofrece los controles adecuados, pero hay que usarlos. La opción **«Autenticar con Microsoft»** viene activada por defecto, y la **autenticación de usuario en las fuentes de conocimiento** hace que el agente muestre a cada persona únicamente el contenido al que esa persona ya tiene acceso.[5] El problema aparece cuando, por comodidad, se publica un agente «sin autenticación» o se le conceden permisos amplios: entonces el agente se convierte en un atajo para saltarse los controles. Las políticas de prevención de pérdida de datos (DLP) permiten precisamente restringir estos conectores y exigir autenticación.[6]

Es exactamente el escenario que el capítulo práctico reproduce en directo: el mismo agente con autenticación de usuario frente a otro publicado «sin autenticación», y cómo una política de datos (DLP) cierra la puerta.

6 Ingeniería social potenciada por IA

EL ATAQUE QUE APUNTA A LA DIRECCIÓN

El último ataque no va contra la máquina, sino contra las personas, y muy especialmente contra quienes tienen capacidad de decisión y de firma. La IA generativa ha eliminado las señales que antes delataban un fraude: los correos de phishing ya no tienen faltas ni traducciones torpes, se redactan en el tono exacto de quien supuestamente escribe y llegan en el momento oportuno. Y ha añadido algo nuevo y perturbador: la falsificación de voz e imagen en tiempo real.

El caso de la ingeniería **Arup** lo resume mejor que ninguna estadística. En 2024, un empleado de su oficina de Hong Kong realizó quince transferencias por un total de unos 25 millones de dólares tras participar en una videollamada en la que todos los demás asistentes, incluido el director financiero, eran deepfakes generados por IA a partir de vídeos públicos de los directivos. El empleado, que había desconfiado del correo inicial, se convenció al «ver» y «oír» a sus superiores.[7] Ninguna herramienta técnica de la empresa falló; falló la confianza en lo que los sentidos daban por cierto.

Aquí la defensa no es tecnológica, sino de procedimiento: verificación por un segundo canal para cualquier orden de pago o

acceso, doble aprobación por encima de un umbral y una desconfianza sana ante la urgencia. Cualquiera de esas medidas habría detenido el fraude de Arup.

La IA no ha inventado el engaño: ha eliminado los errores que antes lo delataban. Verificar por un segundo canal deja de ser paranoia y pasa a ser procedimiento.

Síntesis del capítulo

Seis ataques, una sola raíz: en la era de la IA generativa, quien controla el contenido que el sistema lee puede llegar a controlar lo que el sistema hace. La inyección planta la orden; la exfiltración roba a través del agente; el secuestro le cambia el objetivo; el envenenamiento contamina lo que aprende; el «confused deputy» abusa de sus permisos; y la ingeniería social salta directamente a las personas. No hacen falta hazañas técnicas: basta con palabras bien colocadas.

La defensa, que el capítulo práctico (Capítulo 4) convierte en configuración concreta sobre Copilot Studio, se apoya en unos pocos principios que se repiten: tratar todo contenido externo como no fiable, aplicar el mínimo privilegio, exigir autenticación, limitar la salida de datos con políticas de DLP y mantener siempre a una persona en el lazo de decisión para las acciones que importan. Con el mapa de ataques ya sobre la mesa, pasamos de entender el peligro a saber cerrarlo.

FUENTES

Referencias

1 Aim Security / The Hacker News (junio de 2025). EchoLeak (CVE-2025-32711), CVSS 9,3. thehackernews.com

2 Bargury, M. & Ishay Sharbat, T. / Zenity (Black Hat USA 2024). *Living off Microsoft Copilot*. zenity.io

3 Zenity Labs (2025). *A Copilot Studio Story 2: AIjacking to Full Data Exfiltration*. labs.zenity.io

4 Zenity Labs (2025). «AgentFlayer»: secuestro silencioso de agentes de IA. prnewswire.com

5 Microsoft Learn. *Publish and deploy your agent* — canales y autenticación. learn.microsoft.com

6 Microsoft Learn. *Configure data policies for agents* (DLP). learn.microsoft.com

7 CNN Business (mayo de 2024). Arup, víctima de un fraude con deepfakes por 25 millones de dólares. cnn.com



CAPÍTULO CUATRO

Seguridad práctica en Copilot Studio: crear y compartir agentes

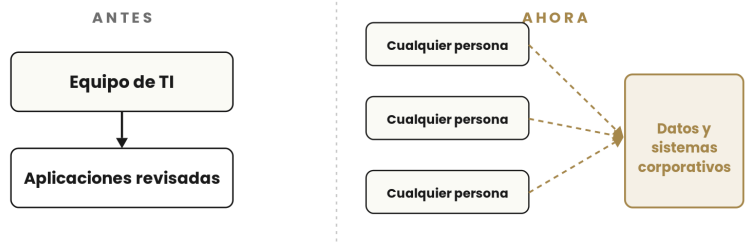
Llegamos al núcleo práctico de la sesión. Hasta aquí hemos entendido los riesgos y los ataques; ahora bajamos a la herramienta concreta con la que hoy se crean agentes en muchas organizaciones, **Microsoft Copilot Studio**, para ver, con casos reales, cómo se rompe y cómo se protege. Todos los casos de este capítulo van acompañados de una demostración en vivo.

El objetivo no es convertir a la dirección en administradores de la plataforma, sino que reconozca las decisiones de configuración que marcan la diferencia entre un agente útil y una brecha esperando a ocurrir. Recorreremos seis casos, del más publicitado a los más sutiles, y cerraremos con las cuatro buenas prácticas que, aplicadas juntas, neutralizan la mayoría de ellos.

• Punto de partida: la creación «no-code» amplía la superficie

La gran promesa de Copilot Studio es que cualquiera puede crear un agente sin escribir código: se describe en lenguaje natural, se le conectan fuentes de datos y herramientas, y se comparte. Esa misma facilidad es su mayor riesgo de seguridad. La capacidad de conectar un sistema a datos corporativos y de publicarlo para otros deja de estar en manos de un equipo técnico que conoce los controles, y pasa a estar en las de cualquier persona con buena voluntad y prisa.

FIGURA 1 · LA CREACIÓN «NO-CODE» AMPLÍA LA SUPERFICIE DE ATAQUE



Pocas manos · camino controlado

Muchas manos crean agentes conectados a datos

El resultado es un cambio de escala en la superficie de ataque. Donde antes había unas pocas aplicaciones revisadas, ahora puede haber decenas o cientos de agentes creados por distintas áreas, cada uno con sus conexiones, sus permisos y su alcance de publicación. La pregunta de gobierno cambia de «¿es seguro este sistema?» a «**¿sabemos siquiera cuántos agentes existen, quién los creó y a qué tienen acceso?**». Con esa idea de fondo, veamos qué puede salir mal.

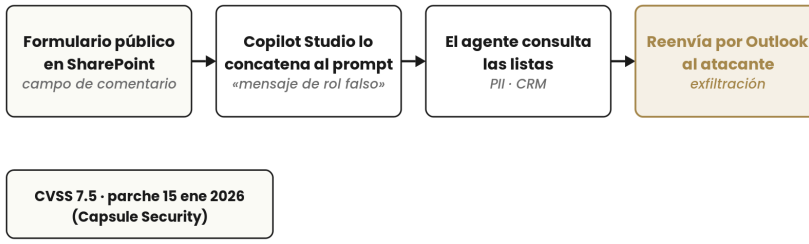
1 Inyección de prompts vía formularios y SharePoint

Es el caso que mejor resume el peligro, y el más reciente. Muchos agentes se conectan a formularios o listas de SharePoint para tareas cotidianas: recoger solicitudes, tickets de soporte o registros de contactos. El problema es que el contenido que un desconocido escribe en un formulario público acaba, sin filtrar, dentro del contexto del agente.

En enero de 2026, los investigadores de Capsule Security divulgaron **«ShareLeak» (CVE-2026-21520)**, una vulnerabilidad de

inyección indirecta en Copilot Studio con severidad 7,5. En agentes activados por el envío de formularios de SharePoint, un atacante rellenaba un campo de comentario con un falso «mensaje de sistema» que Copilot Studio concatenaba directamente al prompt del agente, sin separación entre la entrada no fiable y el modelo. Una vez dentro, la instrucción ordenaba al agente consultar las listas conectadas (datos personales, registros de CRM) y reenviarlos por Outlook a una dirección del atacante.[1] Microsoft publicó un parche, pero los propios investigadores subrayaron lo esencial: parchear un caso no corrige la arquitectura, que sigue tratando la entrada del usuario como si fuera una orden legítima.[1]

FIGURA 2 · INYECCIÓN VÍA FORMULARIO Y EXFILTRACIÓN («SHARELEAK», CVE-2026-21520)



El parche corrige un caso; la arquitectura sigue mezclando la entrada no fiable con el prompt.

EN LA DEMOSTRACIÓN EN VIVO

Montaremos un agente conectado a una lista de SharePoint y enviaremos un formulario cuyo campo de texto contiene una instrucción oculta. Veremos, con datos ficticios cómo el agente la obedece, y después aplicaremos las contramedidas (aislar la entrada, mínimo privilegio y DLP) para cortarla.

2 «Confused deputy»: agentes con las credenciales de su creador

Ya presentamos este patrón en el capítulo anterior; aquí importa su cara práctica. Por defecto, un agente puede acabar ejecutándose con los permisos de la persona que lo creó. Si quien lo construye es alguien con acceso amplio y luego lo comparte con toda la organización, el agente se convierte en un *delegado confundido*: un intermediario de confianza al que cualquiera puede pedir, indirectamente, datos que su propio perfil jamás le permitiría ver. La comodidad de «que funcione para todos» se traduce, sin quererlo, en «que exponga a todos».

EN LA DEMOSTRACIÓN EN VIVO

Compararemos el mismo agente ejecutándose con las credenciales del creador frente a la autenticación de usuario activada, que hace que cada persona vea únicamente el contenido al que ya tiene acceso.

3 Sobreexposición al compartir o publicar

Crear el agente es solo la mitad; la otra mitad es decidir quién puede usarlo, y ahí se cometen los errores más caros. Copilot Studio permite publicar un agente en Teams, incrustarlo en una web o exponerlo mediante un sitio de demostración. Cada canal tiene sus opciones de autenticación, y la más peligrosa es la más cómoda: publicar «sin autenticación», de forma que el agente quede accesible incluso de manera anónima.

Un agente anónimo con acceso a datos internos es una puerta abierta a Internet. La opción «Autenticar con Microsoft» viene activada por defecto precisamente para evitarlo, pero basta un clic mal entendido para desactivarla en aras de la comodidad.[2] En el momento de la publicación, decidir **quién, por qué canal y con qué autenticación** se accede al agente son decisiones de seguridad de primer orden, no un trámite final.

EN LA DEMOSTRACIÓN EN VIVO

Publicaremos un agente en dos modos, con autenticación y en modo anónimo, y veremos con datos ficticios cómo el segundo expone información que el primero protege.

4 Backdoor mediante «Connected Agents»

Este es uno de los más sutiles y recientes. Copilot Studio incorpora una función, **«Connected Agents»**, que permite que un agente use las capacidades, herramientas, conocimiento y temas de otro agente del mismo entorno. Es útil para componer agentes especializados, pero los investigadores de Zenity Labs mostraron su lado oscuro: la función viene **activada por defecto** en los agentes nuevos, y sus invocaciones no dejan rastro en la pestaña de actividad, por lo que un defensor no puede detectar el abuso con la monitorización nativa.[3]

FIGURA 3 · BACKDOOR MEDIANTE «CONNECTED AGENTS»



Activado por defecto · sin rastro en la pestaña de actividad

El ataque es elegante en su simplicidad: alguien crea un agente aparentemente inofensivo y lo conecta a otro con privilegios, por ejemplo, uno capaz de enviar correos desde el dominio oficial de la organización o con acceso a datos sensibles. El agente malicioso hereda así, por la puerta de atrás, las herramientas y el conocimiento del privilegiado. En pruebas de concepto, Zenity comprometió agentes de soporte para lanzar ataques de suplantación a gran escala.[3] La recomendación es directa: **desactivar «Connected Agents»** en todo agente que tenga herramientas sin autenticar o fuentes de conocimiento sensibles antes de publicarlo.

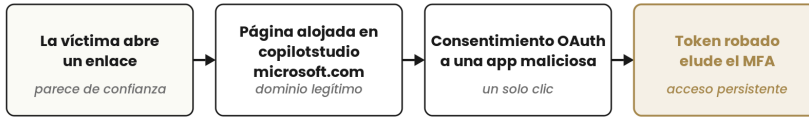
DEMOSTRACIÓN GUIADA

Mostraremos dónde vive el interruptor de «Connected Agents», cómo un agente puede alcanzar las herramientas de otro y por qué la ausencia de registro lo hace tan peligroso. Cerraremos desactivándolo y razonando el inventario de agentes.

5 Phishing de OAuth desde el dominio legítimo («CoPhish»)

Si los anteriores abusan de los datos, este abusa de la confianza. La técnica «CoPhish», descrita por Datadog Security Labs, aprovecha que los agentes de Copilot Studio pueden alojarse en una URL del propio dominio oficial de Microsoft (copilotstudio.microsoft.com). Un atacante crea ahí una página con un botón de inicio de sesión que, en realidad, lanza una solicitud de consentimiento OAuth para una aplicación maliciosa. Como todo sucede bajo un dominio de máxima confianza, la víctima no sospecha; al aceptar, entrega un token de acceso que puede eludir el MFA y conceder acceso persistente.[4]

FIGURA 4 · «COPHISH»: PHISHING DE OAUTH DESDE UN DOMINIO LEGÍTIMO



El ataque abusa de la confianza en la propia plataforma de Microsoft.

La defensa aquí es de identidad y gobierno, más que de la propia herramienta: aplicar una política estricta de consentimiento de aplicaciones, desactivar por defecto que los usuarios puedan crear y consentir aplicaciones, y vigilar de cerca los eventos de consentimiento en Entra ID y la creación de agentes.[4] Es un

recordatorio de que la seguridad de la IA no acaba en el agente: se extiende a cómo se gobierna la identidad alrededor de él.

DEMOSTRACIÓN GUIADA / CONCIENCIACIÓN

Sin robar credenciales de nadie, mostraremos cómo el uso del dominio legítimo hace creíble el engaño y qué controles de consentimiento de aplicaciones en Entra ID lo cierran.

6 Conectores con exceso de permisos y brechas de DLP

El último caso es el más común y el menos vistoso: agentes a los que, por comodidad, se les conceden conectores con más permisos de los necesarios, o que emplean rutas de datos que esquivan las políticas de la organización. Un conector que puede leer y escribir en medio mundo corporativo, o uno que envía datos a un servicio externo no controlado, convierte cualquiera de los ataques anteriores en una fuga real.

Copilot Studio y Power Platform ofrecen **políticas de prevención de pérdida de datos** (DLP) que clasifican los conectores en grupos: de negocio, no de negocio y bloqueados, e impiden que se combinen de forma peligrosa; desde 2025, su aplicación es obligatoria para todos los agentes del entorno.[5] La brecha aparece cuando estas políticas no existen, están en modo permisivo o no contemplan los conectores que de verdad se usan. El mínimo privilegio en cada conector no es una recomendación estética: es lo que limita el alcance del daño cuando algo falla.

EN LA DEMOSTRACIÓN EN VIVO

Veremos un agente con un conector de salida sin restringir intentando enviar datos fuera, y cómo una política DLP que bloquea ese conector detiene el flujo.

Casi todos estos casos no explotan un fallo del producto, sino una decisión de configuración cómoda. La seguridad de un agente se decide al crearlo y al compartirlo.

Buenas prácticas de configuración segura

Los seis casos anteriores se defienden con cuatro principios que conviene aplicar siempre juntos. Ninguno es exótico; todos están al alcance de la configuración de la plataforma.

1 · MÍNIMO PRIVILEGIO EN CONECTORES Y DATOS

Cada agente debe tener acceso solo a lo estrictamente necesario para su tarea, y ni un dato más. Conectores con el alcance más reducido posible, cuentas de servicio con permisos mínimos y ninguna conexión «por si acaso». Cuanto menos pueda tocar un agente, menos podrá filtrar o ejecutar si es manipulado.

2 · AUTENTICACIÓN OBLIGATORIA Y ALCANCE DE PUBLICACIÓN CONTROLADO

Mantener «Autenticar con Microsoft» y la autenticación de usuario en las fuentes de conocimiento, de modo que cada persona vea

únicamente lo suyo. Nada de canales anónimos con acceso a datos internos. Cada publicación debe responder conscientemente a quién, por qué canal y con qué autenticación.

3 · POLÍTICAS DE PREVENCIÓN DE PÉRDIDA DE DATOS (DLP)

Definir y aplicar políticas DLP que clasifiquen los conectores y bloqueen las combinaciones y las salidas peligrosas, incluida la conexión de chat sin autenticación de Entra ID. Revisarlas cuando aparezcan nuevos conectores, y no dejarlas nunca en un modo meramente informativo cuando protejan datos sensibles.

4 · MONITORIZACIÓN Y AUDITORÍA (MICROSOFT PURVIEW / DEFENDER)

Lo que no se ve no se puede gobernar. Microsoft Purview ofrece visibilidad del uso de la IA, detecta información sensible en prompts y respuestas, evalúa riesgos de sobreexposición y habilita la auditoría de las interacciones con Copilot y los agentes; Defender complementa la detección de amenazas.[6] Inventariar los agentes, saber quién los crea y vigilar su comportamiento es lo que convierte todo lo anterior en un sistema gobernado y no en una foto fija.

Síntesis del capítulo

La lección de este capítulo, el más práctico de todos, es tranquilizadora: la inmensa mayoría de los incidentes que hemos visto no nacen de una tecnología imposible de proteger, sino de decisiones de configuración cómodas tomadas sin ver el riesgo. Aislar la entrada no fiable, aplicar el mínimo privilegio, exigir autenticación, controlar la publicación, desactivar lo que no se

necesita (como «Connected Agents»), gobernar el consentimiento de aplicaciones y vigilar con DLP y Purview cierran, juntos, casi todas las puertas. Falta la última pieza: las reglas, los responsables y los procesos que sostienen estas decisiones en el tiempo. De eso trata la gobernanza, el capítulo siguiente.

FUENTES

Referencias

1 Capsule Security (enero de 2026). *ShareLeak: Taking the Wheel of Microsoft's Copilot Studio (CVE-2026-21520)*. capsulesecurity.io

2 Microsoft Learn. *Publish and deploy your agent*. learn.microsoft.com

3 Zenity Labs (2025). *Connected Agents: The hidden agentic puppeteer*. labs.zenity.io

4 Datadog Security Labs (octubre de 2025). *CoPhish*. securitylabs.datadoghq.com

5 Microsoft Learn. *Configure data policies for agents*. learn.microsoft.com

6 Microsoft Learn. *Use Microsoft Purview to manage data security & compliance for Copilot Studio*. learn.microsoft.com



CAPÍTULO CINCO

Gobernanza, cumplimiento y responsabilidad directiva

Los capítulos anteriores nos han dado tres cosas: entender los riesgos, conocer los ataques y saber configurar la herramienta de forma segura. Pero falta la pieza que sostiene a las demás. Un control técnico sin gobernanza *caduca*: se desactiva por comodidad, se olvida al cambiar de responsable, o simplemente nadie lo revisa. La gobernanza es lo que convierte la seguridad en algo duradero y, sobre todo, lo que reparte con claridad quién responde de qué.

Este capítulo es el más directivo de todos, porque su materia no es técnica: son reglas, funciones y procesos. No se trata de burocracia por la burocracia, sino de la diferencia entre una organización que apaga fuegos y otra que gobierna su riesgo. Lo veremos en cinco piezas: qué se puede introducir en una IA, cómo se gestiona el ciclo de vida de un agente, qué obliga la ley, quién responde de las decisiones y qué hacer cuando algo falla.

1 Política de uso aceptable y clasificación de datos

El artefacto de gobernanza más rentable no es un sistema caro: es una política de uso aceptable clara y breve, acompañada de una clasificación de datos. Responde a la pregunta que todo el mundo se hace y casi nadie tiene contestada: ¿qué puedo y qué no puedo escribir en una herramienta de IA?

La base es clasificar la información en unos pocos niveles, por ejemplo: **pública, interna, confidencial y de categoría especial**, y decir, para cada nivel, qué herramientas están permitidas. En una universidad esto es especialmente sensible: datos personales de estudiantes, expedientes, evaluaciones, datos de salud o de

investigación, información sujeta a confidencialidad. La regla debe ser tan concreta que un profesor con prisa la entienda en diez segundos: «los datos personales o confidenciales no se pegan nunca en herramientas de IA de consumo; para eso está la versión corporativa puesta a disposición del personal».

Como vimos en el primer capítulo, la fuga se produce en un prompt, casi siempre sin mala intención. La política no sirve para castigar, sino para que la opción correcta sea también la fácil: si la organización ofrece una herramienta segura y dice con claridad qué va en cada sitio, la mayoría de las fugas desaparecen antes de ocurrir.

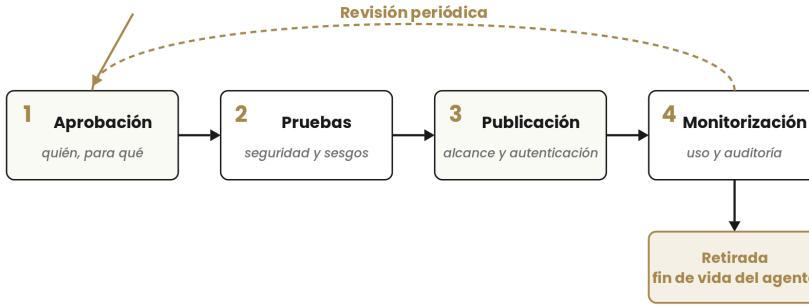
DECISIÓN DE LA DIRECCIÓN

Aprobar una política de uso aceptable de una página, clasificar los datos en niveles y designar la herramienta corporativa sancionada. Sin esas tres cosas, cada persona improvisa su propia política y la de la comodidad casi siempre gana.

2 Ciclo de vida del agente

Un agente no es un documento que se crea y se olvida: es un sistema vivo, conectado a datos y capaz de actuar. Por eso necesita un ciclo de vida gobernado, igual que cualquier otro activo crítico. La facilidad de crearlos «sin código» hace este punto aún más urgente: si cualquiera puede montar uno en diez minutos, la organización necesita saber cuántos existen, quién los creó y a qué acceden.

FIGURA 1 · CICLO DE VIDA DEL AGENTE



Ningún agente debería crearse, conectarse a datos ni compartirse sin pasar por este ciclo, y cada agente debe tener un responsable asignado.

El ciclo tiene cinco fases. La **aprobación** define quién puede crear un agente y para qué, y le asigna un responsable. Las **pruebas** comprueban seguridad y sesgos antes de exponerlo. La **publicación** decide conscientemente el alcance y la autenticación, como vimos en el capítulo práctico. La **monitorización** vigila el uso y deja rastro auditable. Y la **revisión periódica** reevalúa cada agente cada cierto tiempo, porque los datos, los permisos y las personas cambian.

El momento que más se olvida es el último: la retirada. Los agentes abandonados, cuyo creador cambió de puesto o que dejaron de usarse, son un riesgo silencioso: siguen conectados a datos, siguen publicados, y ya no los vigila nadie. Retirar lo que no se usa es una medida de higiene tan básica como cerrar una cuenta de un empleado que se marcha.

DECISIÓN DE LA DIRECCIÓN

Establecer un control de aprobación antes de conectar un agente a datos o publicarlo, un responsable por agente, revisiones periódicas y la retirada obligatoria de los agentes huérfanos. Y, como base de todo: un inventario de qué agentes existen.

3 Marco normativo: RGPD, EU AI Act y deberes de confidencialidad

La responsabilidad directiva sobre la IA no es solo una buena práctica: es una obligación legal creciente, y el desconocimiento no exime de ella. Para una universidad pública, tres marcos se solapan y conviene conocerlos con algo más de detalle, porque determinan qué se puede hacer, con qué garantías y quién responde.

EL RGPD: LA BASE QUE LA IA TENSIONA

El Reglamento General de Protección de Datos se aplica siempre que una herramienta de IA trate datos personales, y en una universidad casi todo lo son: estudiantes, personal, investigación, etc. La IA no cambia las reglas, pero pone a prueba varios de sus principios.

Cuatro exigencias merecen atención especial. Primero, la **base de licitud**: no se puede tratar un dato personal en una IA «porque sí»; hace falta un fundamento válido (consentimiento, interés público, obligación legal…). Segundo, la **minimización de datos**:

introducir solo lo estrictamente necesario, nunca «el expediente entero por comodidad». Tercero, los límites a las **decisiones automatizadas** del artículo 22: ninguna decisión que produzca efectos jurídicos o significativos sobre una persona, como pueden ser admitir a un estudiante, calificar o seleccionar personal, puede adoptarse únicamente por una máquina, sin intervención humana y sin que el afectado pueda conocer la lógica y reclamar. Cuarto, la **evaluación de impacto (EIPD/DPIA)**: los tratamientos de alto riesgo exigen analizar y documentar sus riesgos antes de ponerlos en marcha (art. 35).

A ello se añaden dos cuestiones que la IA vuelve muy concretas. Las **transferencias internacionales**: muchas herramientas procesan los datos en servidores fuera de la Unión Europea, lo que requiere garantías adicionales; por eso importa tanto elegir la versión y la región del servicio. Y la **notificación de brechas**: si se produce una fuga de datos personales, hay que comunicarla a la autoridad, siendo en España la AEPD (Agencia Española de Protección de Datos), en un plazo de 72 horas.[1]

EL REGLAMENTO EUROPEO DE IA (EU AI ACT): NIVELES DE RIESGO Y CALENDARIO

El Reglamento Europeo de IA no regula «la IA» en bloque, sino según su riesgo, en cuatro niveles. Entenderlos ayuda a situar cada uso.

- **Riesgo inaceptable (prohibido):** usos vetados, como la puntuación social o ciertas manipulaciones. En vigor desde febrero de 2025, con sanciones de hasta 35 millones de euros o el 7 % de la facturación mundial.

- **Alto riesgo:** sistemas con impacto significativo en las personas. El Anexo III incluye expresamente la educación (admisión, evaluación de estudiantes, asignación a centros), además del empleo y el crédito.
- **Riesgo limitado (transparencia):** chatbots y contenido generado por IA, que deben identificarse como tales. Exigible desde el 2 de agosto de 2026.
- **Riesgo mínimo:** la mayoría de usos cotidianos, sin obligaciones específicas.

El calendario es clave, y ha cambiado. Las obligaciones de **transparencia** que incluyen, por ejemplo, avisar de que se está hablando con una IA y etiquetar imágenes o vídeos sintéticos, son exigibles desde agosto de 2026.[2] Sin embargo, las obligaciones más duras, las de los **sistemas de alto riesgo**, fueron aplazadas por el llamado «Digital Omnibus» a **diciembre de 2027** (y a agosto de 2028 para la IA embebida en productos regulados).[2] Esas obligaciones son exigentes: gestión de riesgos, gobernanza de los datos de entrenamiento, documentación técnica, registro de la actividad, transparencia, **supervisión humana** (art. 14) y niveles adecuados de precisión, robustez y ciberseguridad.

Para una universidad esto es doblemente relevante: si emplea IA para **admitir o evaluar estudiantes**, estará ante un sistema de **alto riesgo**, con todas esas obligaciones. El aplazamiento a diciembre de 2027 no es una exención, sino una *pista de despegue*: el momento de inventariar esos usos, documentarlos y prepararlos es ahora, no en 2027.

ESPAÑA: AEPD, AESIA, ENS Y EL DEBER DE SECRETO

En el plano nacional confluyen varias piezas que una universidad pública debe tener presentes. La **AEPD** supervisa el cumplimiento del RGPD.[3] La **AESIA** (Agencia Española de Supervisión de la Inteligencia Artificial) es la autoridad designada para vigilar el cumplimiento del Reglamento de IA en España, y será quien supervise y, en su caso, sancione.[4] Y, de forma muy concreta, el **Esquema Nacional de Seguridad (ENS)** (Real Decreto 311/2022) es de cumplimiento **obligatorio** para todo el sector público, incluidas las universidades públicas, y actúa como vehículo técnico para cumplir el RGPD en el sector público. El ENS clasifica los sistemas por niveles (básico, medio y alto) y exige medidas de seguridad proporcionadas a cada uno.[5]

A todo ello se suman los deberes propios de confidencialidad y secreto sobre los datos de estudiantes, del personal y de la investigación, incluida la información sujeta a propiedad industrial o a acuerdos de confidencialidad con terceros. Introducir cualquiera de esos datos en una IA no gobernada puede vulnerar, a la vez, el RGPD, el ENS y un deber de secreto contractual.

PARA LA UNIVERSIDAD, EN CONCRETO

Tres consecuencias prácticas: (1) cualquier uso de IA para admisión o evaluación de estudiantes es alto riesgo y debe prepararse ya; (2) el ENS es obligatorio y aplica a los sistemas donde viva la IA; (3) los tratamientos de datos personales relevantes necesitan una evaluación de impacto (EIPD) y una base de licitud clara, con el delegado de protección de datos implicado desde el principio.

El aplazamiento de las reglas de alto riesgo es una pista de despegue, no una exención. La dirección responde del cumplimiento; el desconocimiento no exime.

4 Rendición de cuentas y supervisión humana

Detrás de toda la técnica y la norma hay una idea sencilla y no negociable: la responsabilidad última de una decisión es siempre humana. La IA propone, acelera, redacta; pero quien decide, y quien responde, es una persona. Delegar una decisión en una máquina no diluye la responsabilidad: la concentra en quien decidió delegarla.

Este principio, el «*human in the loop*» (persona en el bucle de decisión), no es sólo ético: el propio Reglamento de IA lo exige como supervisión humana para los sistemas de alto riesgo. Y los tribunales ya lo han aplicado: recordemos el caso **Air Canada** del primer capítulo, donde la organización fue considerada responsable de lo que dijo su asistente. No vale escudarse en que «lo dijo la IA»: ante un tercero, la IA es la organización.

Para la dirección, esto se traduce en una regla operativa: toda decisión relevante que se apoye en IA, ya sea sobre personas, dinero o derechos, debe tener un responsable humano identificable que la revise y la asuma. Cuanto mayor sea la

consecuencia, más estricta debe ser esa supervisión. La rendición de cuentas no se automatiza.

5 Respuesta ante incidentes

Por muy bien que se gobierne, algo acabará fallando: una fuga, un agente mal configurado, una decisión sesgada, un ataque. La diferencia entre un susto y una crisis está en haber decidido, antes de que ocurra, qué se hace y a quién se avisa. Improvisar en caliente es la peor de las opciones.

FIGURA 2 · RESPUESTA ANTE INCIDENTES



El plan se prepara antes del incidente: quién decide, a quién se avisa y en qué plazo.

El plan tiene cinco pasos: **detectar** qué ha pasado, **contener** para limitar el daño, **notificar** a quien corresponda (si hay datos personales afectados, a la AEPD en un plazo de 72 horas),[1] **investigar** el alcance y la causa, y **aprender** para que no vuelva a ocurrir. Cada paso necesita un responsable con nombre y un canal claro: ¿a quién llama el profesor que se da cuenta de que ha filtrado datos? Si la respuesta no está escrita, no existe.

Y una lección que este propio curso nos ha regalado: **cerrar la puerta no es inmediato**. Entre que se detecta un problema y que la mitigación surte efecto hay una ventana de exposición en la que el

daño puede seguir produciéndose. Por eso conviene asumir, ante un incidente, que el dato pudo salir, y actuar en consecuencia en vez de confiar en que «ya lo hemos apagado».

DECISIÓN DE LA DIRECCIÓN

Tener un plan de respuesta escrito antes del incidente: quién detecta, quién decide, a quién se notifica y en qué plazo. Un incidente bien gestionado protege a la institución; uno improvisado la expone dos veces.

Síntesis del capítulo

La gobernanza es el hilo que cose todo lo anterior. Entendimos los riesgos, conocimos los ataques, aprendimos a configurar la herramienta con seguridad y, ahora, hemos puesto las reglas, las funciones y los procesos que hacen que todo eso perdure y tenga un responsable. Cinco decisiones resumen el capítulo: una política de uso y clasificación de datos, un ciclo de vida para los agentes, el cumplimiento del marco legal, la supervisión humana de las decisiones y un plan de respuesta ante incidentes. Con esto, la organización deja de depender de la buena voluntad individual y pasa a gobernar su riesgo. Solo queda convertirlo en acción: el capítulo final es la lista de verificación y el plan concreto para la dirección.

FUENTES

Referencias

1 Reglamento (UE) 2016/679 (RGPD): notificación de brechas en 72 h (art. 33), decisiones automatizadas (art. 22) y evaluación de impacto (art. 35). eur-lex.europa.eu

2 Reglamento Europeo de IA (EU AI Act) y «Digital Omnibus» (2026): prohibidos desde feb. 2025; transparencia (art. 50) desde ago. 2026; alto riesgo (Anexo III, educación incluida) aplazado a Dic. 2027. artificialintelligenceact.eu

3 Agencia Española de Protección de Datos (AEPD). aepd.es

4 AESIA — Agencia Española de Supervisión de la Inteligencia Artificial. aesia.digital.gob.es

5 Esquema Nacional de Seguridad (ENS), Real Decreto 311/2022: obligatorio para el sector público, incluidas las universidades públicas. ccn-cert.cni.es



CAPÍTULO SEIS

Cierre, checklist y plan de acción

Hemos recorrido mucho camino: entendimos las oportunidades y los riesgos de la IA generativa, vimos cómo se ataca, aprendimos a configurarla con seguridad y pusimos las reglas que hacen todo eso duradero. Este último capítulo tiene un único propósito: que nada de lo anterior se quede en apuntes. Lo convertimos en una lista de verificación que la dirección puede empezar a usar mañana, un plan por fases y respuestas a las preguntas que probablemente estén rondando la sala.

1 Lista de verificación para la dirección

No hay que hacerlo todo a la vez ni hay que ser experto en tecnología. Basta con ir marcando estas casillas, por bloques. Si al terminar la sesión solo se recuerda una página del curso, que sea esta.

DATOS Y USO ACEPTABLE

- ☐ Aprobar una política de uso aceptable de una sola página.
- ☐ Clasificar los datos en niveles: pública, interna, confidencial y de categoría especial.
- ☐ Designar la herramienta de IA corporativa sancionada, con tratamiento de datos garantizado.
- ☐ Regla clara y conocida: los datos personales o confidenciales no se introducen nunca en IA de consumo.

AGENTES: CREACIÓN Y CICLO DE VIDA

- ☐ Inventariar los agentes que ya existen: quién los creó, a qué datos acceden y dónde están publicados.
- ☐ Establecer una puerta de aprobación antes de conectar un agente a datos o publicarlo.
- ☐ Asignar un responsable a cada agente.
- ☐ Programar revisiones periódicas y retirar los agentes huérfanos o en desuso.

CONFIGURACIÓN SEGURA

- ☐ Autenticación obligatoria; ningún agente anónimo con acceso a datos internos.
- ☐ Mínimo privilegio en conectores y acceso a datos.
- ☐ Políticas de prevención de pérdida de datos (DLP) aplicadas, no en modo permisivo con datos sensibles.
- ☐ Desactivar lo que no se necesita (por ejemplo, «Connected Agents»).
- ☐ Monitorización y auditoría con Microsoft Purview / Defender.

GOBERNANZA Y CUMPLIMIENTO

- ☐ Tener presente el marco legal: RGPD, Reglamento Europeo de IA y ENS (universidad pública).
- ☐ Evaluación de impacto (EIPD) y base de licitud para los tratamientos relevantes, con el DPD implicado.
- ☐ Supervisión humana en toda decisión relevante sobre personas, dinero o derechos («human in the loop»).

- ☐ Plan de respuesta a incidentes escrito: quién detecta, quién decide, a quién se notifica y en qué plazo (AEPD, 72 h).
- ☐ Controlar el consentimiento de aplicaciones en la identidad (Entra ID).

PERSONAS Y CULTURA

- ☐ Formar a la dirección y al personal: son el objetivo prioritario de la ingeniería social.
- ☐ Verificación por un segundo canal para pagos, accesos o cambios sensibles.
- ☐ Fomentar el uso seguro y ofrecer alternativas, en lugar de limitarse a prohibir.

2 Próximos pasos y recomendaciones

La checklist puede abrumar si se mira entera. La clave es el orden: unos pocos pasos, bien secuenciados, sostienen casi toda la seguridad. Este es un plan realista para empezar sin bloquearse.

FIGURA 1 · PLAN DE ACCIÓN PARA LA DIRECCIÓN



De entender el riesgo a gobernarlo: pequeños pasos, en orden, sostienen toda la seguridad.

Tres recomendaciones de fondo cierran el recorrido. La primera: empezar por lo barato y lo rápido (nombrar un responsable, inventariar y avisar) tiene un retorno enorme y no cuesta prácticamente nada. La segunda: gobernar no es prohibir; la organización que ofrece una alternativa segura y unas reglas claras aprovecha la IA mejor y con menos sustos que la que la persigue. Y la tercera, la que vertebra todo el curso: la tecnología cambia cada semana, pero los principios no. Mínimo privilegio, autenticación, supervisión humana y tratar el contenido externo como no fiable seguirán siendo válidos cuando las herramientas de hoy hayan cambiado de nombre.

Cierre

Si hubiera que resumir todo el curso en una sola idea, sería esta: la IA generativa ya está dentro de la organización, y la única elección real no es usarla o no, sino hacerlo de forma gobernada o dejarla a la sombra. No hace falta miedo ni hace falta euforia; hace falta criterio. Con unas pocas decisiones de dirección claras, ordenadas y sostenidas en el tiempo es perfectamente posible aprovechar todo lo que la IA ofrece sin poner en riesgo lo que la institución debe proteger. Ese equilibrio, y no otra cosa, es lo que significa gobernar la inteligencia artificial.

La elección no es usar la IA o no. Es gobernarla o dejarla a la sombra. Y gobernarla está al alcance de la dirección.

Gracias.

Miguel Angel Hernandez Ruiz · @whitetie
https://www.linkedin.com/in/whitetie/

Sobre el autor

Más allá de dejar aquí mi huella con una retahíla de titulaciones y certificaciones, os voy a decir algo realmente sobre mí. He estado a punto de rendirme en muchas ocasiones, pero nunca lo he hecho. Vengo de una familia muy humilde y estudié becado porque el peso de una carrera era mucho a espaldas de unos padres que trabajaban en el campo.

No acabé de encontrar la forma de aprobar los exámenes en la universidad en mi primer año, a pesar de todos mis esfuerzos, pero volví a la carga en el segundo y aprobé 13 asignaturas.

No me resigné a quedarme en un país que me daba oportunidades pero no reconocimiento, así que acepté una propuesta para trabajar en la segunda compañía más grande del mundo por número de empleados (IBM). En un momento de mi vida en el que lo tenía absolutamente todo en mi país, puse una pequeña parte de mi ropa en una maleta y me fuí a Irlanda, un país que no conocía y con un idioma que no dominaba del todo, para luchar por un futuro que hoy es presente.

He trabajado para toda clase de empresas, públicas y privadas, en 5 países diferentes y de todas las dimensiones. También he sufrido situaciones muy desagradables cuando no les concedían los proyectos y te “invitaban”, de forma poco sutil, a que buscaras la puerta de salida. No lo hice.

Sigo aprendiendo cada día y probablemente, siempre lo haré. La ciberseguridad es mi pasión aún después de 20 años de dedicación, ahora también lo es la Inteligencia Artificial y cómo aplicarla a este campo laboral que tanto me ha dado durante mis 20 años de carrera profesional.



© 2026 · Miguel Ángel Hernández Ruiz

Esta obra está publicada bajo una licencia Creative Commons Reconocimiento-NoComercial-CompartirIgual 4.0 Internacional (CC BY-NC-SA 4.0).

Eres libre de compartir y adaptar el material, con atribución y para usos no comerciales; las adaptaciones deben compartirse bajo la misma licencia.

creativecommons.org/licenses/by-nc-sa/4.0/deed.es