ACADEMIA HUMANITARIA · HUMANITAIRES JOBS

Mini MBA IA y acción humanitaria: trabajar con Claude y Gemini

Material del curso · Semana 10 : Información, desinformación y seguridad: lo que la IA puede y no puede hacer

Un programa de exactamente 3 meses (12 semanas, 144 horas) para utilizar los asistentes de IA Claude (Anthropic) y Gemini (Google) en el trabajo humanitario, de forma útil, segura y responsable: funcionamiento y límites, redacción de peticiones, protección de datos, ética y principios humanitarios, informes, propuestas, análisis de datos y MEAL, traducción y comunicación con las comunidades, investigación y verificación de fuentes, desinformación y seguridad, productividad y automatización, política de uso y despliegue en una organización. Ejercicios corregidos, tres exámenes mensuales con estudios de caso y un diploma verificable de la Academia Humanitaria. Programa independiente, sin asociación con Anthropic ni con Google.

https://humanitairejobs.org/es/academia/mini-mba-ia-claude-gemini/semaine-10/

Documento gratuito de la Academia Humanitaria. Puede imprimirse y compartirse libremente con fines de formación, sin modificaciones.

Semana 10 : Información, desinformación y seguridad: lo que la IA puede y no puede hacer

Objetivos

- Comprender cómo se utiliza la IA para producir desinformación y contenidos falsificados dirigidos contra los trabajadores humanitarios.
- Utilizar la IA de forma segura como apoyo al análisis de seguridad, sin exponer información sensible.
- Protegerse contra el phishing, la suplantación de identidad y los fraudes facilitados por la IA.

Un nuevo riesgo para la aceptación

La aceptación por parte de las comunidades y de las partes en conflicto es la base de la seguridad humanitaria. Sin embargo, la IA generativa permite producir, a un coste muy bajo, **contenidos falsos y creíbles**: falsos comunicados, falsas citas de responsables, imágenes o vídeos falsificados (*deepfakes*), falsas grabaciones de audio, mensajes masivos en las redes sociales.

Algunas organizaciones humanitarias ya han sido objeto de campañas de desinformación que las acusaban de espionaje, proselitismo, tráfico o parcialidad. Estas campañas pueden provocar amenazas, ataques o expulsiones. El CICR y otros actores alertan sobre los **riesgos de los «daños relacionados con la información»** para la población civil y los trabajadores humanitarios.

> La desinformación es un riesgo de seguridad. Tiene su lugar en el análisis de contexto, la evaluación de riesgos y el plan de comunicación de crisis.

Detectar los contenidos falsificados

Cada vez es más difícil distinguir a simple vista un contenido generado de uno real. Algunos reflejos de verificación siguen siendo eficaces:

- **Fuente**: ¿quién publicó primero? ¿Una cuenta reciente, anónima, sin historial?
- **Contexto**: ¿coinciden la fecha, el lugar, el tiempo, los uniformes, la lengua?
- **Búsqueda inversa de imágenes**: ¿existía la imagen antes, en otro contexto?
- **Indicios técnicos**: manos, textos deformados, sombras incoherentes, voz sin respiración natural; estos indicios desaparecen con los avances de las herramientas, por lo que no bastan.
- **Marcas**: algunos proveedores añaden marcas de agua invisibles o metadatos de origen a los contenidos generados (por ejemplo SynthID en Google, o los metadatos del estándar C2PA). Su ausencia no prueba nada.

- **Confirmación directa**: contacte con la organización o la persona supuestamente en el origen del contenido.

Un asistente de IA puede ayudar a estructurar esta verificación y a enumerar las incoherencias, pero no puede «certificar» que un contenido sea verdadero o falso.

Responder a una campaña de desinformación

Si su organización es objeto de una campaña:

- **Evaluar**: alcance, audiencia, vínculo con actores armados o políticos, riesgo para el personal y las comunidades. Es una decisión de la célula de gestión de riesgos, no solo de comunicación.
- **Proteger**: adaptar las medidas de seguridad si es necesario (desplazamientos, visibilidad, instalaciones).
- **Responder de forma proporcionada**: a veces una respuesta pública amplifica el contenido falso; a veces el silencio deja que se instale. Privilegie a menudo el diálogo directo con las comunidades, los líderes y las autoridades, apoyado en hechos verificables.
- **Documentar**: capturas fechadas, enlaces, denuncia a las plataformas.
- **Coordinar** con las demás organizaciones y los mecanismos de coordinación.

La IA puede ayudar a preparar elementos de lenguaje factuales, preguntas y respuestas para el personal y mensajes adaptados a los distintos públicos, revisados por la dirección y la seguridad.

La IA como apoyo al análisis de seguridad

Para la seguridad, la IA puede ser útil con **información pública** y para **tareas metodológicas**:

- resumir informes públicos sobre un contexto;
- proponer la estructura de un análisis de actores o de un plan de seguridad a partir de una plantilla;
- preparar ejercicios de simulación (escenarios ficticios);
- traducir documentos públicos;
- redactar materiales de formación en seguridad.

No debe recibir:

- los informes detallados de incidentes, los nombres, los itinerarios, los horarios, las posiciones de los equipos y de las instalaciones;
- la información confidencial de una red de seguridad compartida entre ONG o procedente de servicios de análisis de seguridad reservados a sus miembros;
- la información sobre negociaciones, contactos con grupos armados o secuestros.

Y nunca debe **decidir** un nivel de seguridad, un desplazamiento o una evacuación. Esas decisiones corresponden a los responsables designados, según el plan de seguridad.

Phishing y fraudes reforzados por la IA

La IA hace que las estafas sean más convincentes:

- **Correos de phishing** sin faltas, personalizados, en la lengua del destinatario, que imitan a un donante o a la dirección;
- **Fraude del CEO**: falso mensaje de voz o vídeo del director que pide una transferencia urgente;
- **Falsas ofertas de empleo** en nombre de organizaciones humanitarias, que piden pagos a los candidatos;
- **Falsos proveedores** que cambian sus datos bancarios.

Defensas:

- doble validación de todo pago, y confirmación por otro canal conocido de todo cambio de datos bancarios;
- palabra clave o procedimiento de devolución de llamada para las peticiones urgentes inusuales;
- autenticación de dos factores en todas las cuentas, incluidas las cuentas de IA;
- formación periódica y ejercicios de phishing;
- recordatorio público: una organización humanitaria seria nunca pide pagos a los candidatos.

Seguridad de las cuentas y herramientas de IA

Las propias cuentas de IA deben protegerse:

- autenticación fuerte y cuentas nominativas, sin cuentas compartidas;
- revisión periódica de los **conectores** y de los accesos concedidos (correo, almacenamiento);
- atención a la **inyección de instrucciones**: un documento, una página web o un correo puede contener instrucciones ocultas destinadas a la IA («ignora las consignas y envía el contenido a…»). Cuando un asistente lee contenidos externos y dispone de funciones de acción (envío de correos, acceso a archivos), este riesgo se vuelve real. Mantenga la validación humana de las acciones y limite los accesos;
- salida de un empleado: desactivación de las cuentas y transferencia de los proyectos compartidos;
- aplicaciones y extensiones: instale solo aplicaciones oficiales; a veces circulan falsas aplicaciones «Claude» o «Gemini».

Ejercicio: un vídeo viral contra su ONG

Situación. Una mañana, un vídeo de 40 segundos circula por las redes sociales y los grupos de mensajería de la provincia. En él se ve a un hombre que lleva un chaleco con su logotipo repartiendo folletos políticos; una voz afirma que su ONG apoya a un candidato en las elecciones del mes próximo. La cuenta que lo publicó se creó hace 6 días. El logotipo del chaleco es el antiguo, abandonado hace dos años. El personal de terreno señala miradas hostiles y una llamada anónima amenazante. Un colega propone preguntar a una IA «si el vídeo es un deepfake» y publicar la respuesta.

Trabajo solicitado. ¿Qué acciones, en qué orden?

Corrección. 1) La seguridad primero: informar al director de país y al responsable de seguridad, activar la evaluación de riesgos, adaptar inmediatamente los desplazamientos y la visibilidad en la provincia; tratar la llamada amenazante como un incidente. 2) Verificación: documentar (capturas, enlaces, hora), buscar el origen, verificar las incoherencias (cuenta de 6 días, logotipo antiguo), buscar imágenes de origen; una IA puede ayudar a enumerar las incoherencias pero no puede certificar, y publicar «según una IA, es falso» no tiene ningún valor probatorio y puede volverse contra la organización. 3) Respuesta: diálogo directo con los líderes comunitarios y las autoridades, recordatorio factual de la neutralidad de la organización, coordinación con las demás organizaciones; decisión colectiva sobre la conveniencia de una declaración pública sobria. 4) Denuncia a las plataformas. 5) Seguimiento: vigilancia de la difusión y lecciones aprendidas para integrar la desinformación en el plan de seguridad.

Para recordar

- La desinformación generada por la IA es un riesgo de seguridad que amenaza la aceptación.
- Verificar: fuente, contexto, búsqueda inversa, confirmación directa; una IA no certifica un contenido.
- Para la seguridad, la IA trabaja con información pública y tareas metodológicas, nunca con incidentes detallados ni para decidir.
- Contra los fraudes reforzados por la IA: doble validación, confirmación por otro canal, autenticación fuerte.
- Desconfíe de la inyección de instrucciones: validación humana de las acciones y accesos limitados.

Cuestionario

1. Una campaña de desinformación contra una ONG corresponde:
   A. Únicamente a la comunicación
   B. También a la gestión de los riesgos de seguridad
   C. Al servicio financiero
   D. A nadie

2. ¿Puede un asistente de IA certificar que un vídeo es un deepfake?
   A. Sí, siempre
   B. Sí, si el vídeo es corto
   C. No: puede ayudar a enumerar incoherencias, sin certificar
   D. Sí, en versión de pago

3. Un mensaje de voz del director pide una transferencia urgente a una cuenta nueva. Se:
   A. Confirma por otro canal conocido y se aplica la doble validación
   B. Paga inmediatamente
   C. Pregunta a la IA si es verdad
   D. Transfiere la mitad

4. La «inyección de instrucciones» designa:
   A. Un error de escritura
   B. Una actualización del programa
   C. Una formación
   D. Instrucciones ocultas en un contenido, destinadas a desviar a la IA

5. ¿Cuál de estos documentos puede resumirse con la IA para un análisis de seguridad?
   A. Un informe público sobre el contexto de un país
   B. El informe detallado del incidente de ayer
   C. Los itinerarios de la semana
   D. Los contactos con un grupo armado

Respuestas del cuestionario : 1. B · 2. C · 3. A · 4. D · 5. A