ACADEMIA HUMANITARIA · HUMANITAIRES JOBS

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

Material del curso · Semana 4 : Ética de la IA y principios humanitarios: no hacer daño, sesgos, transparencia

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-4/

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

-

Semana 4 : Ética de la IA y principios humanitarios: no hacer daño, sesgos, transparencia

Objetivos

- Relacionar el uso de la IA con los principios humanitarios (humanidad, neutralidad, imparcialidad, independencia) y con el principio de no hacer daño.
- Detectar los sesgos de una respuesta de IA y sus posibles efectos en las personas afectadas.
- Aplicar reglas claras de transparencia, responsabilidad humana y respeto de los derechos de autor.

Los principios humanitarios ante la IA

Los cuatro principios humanitarios se aplican también a las herramientas digitales:

- **Humanidad**: la herramienta debe servir a las personas afectadas, no solo a la eficiencia interna. Ahorrar tiempo en un informe es útil si ese tiempo vuelve a los programas.
- **Imparcialidad**: una IA no debe introducir discriminación en el acceso a la ayuda (por ejemplo clasificando hogares según criterios sesgados).
- **Neutralidad**: un texto generado puede retomar, sin que se note, el vocabulario de una parte en conflicto. Revise los nombres de lugares, grupos y acontecimientos.
- **Independencia**: depender por completo de un proveedor para una función crítica es un riesgo; conserve la capacidad de trabajar sin la herramienta.

El principio de **no hacer daño** (*do no harm*) obliga a preguntarse antes de cada nuevo uso: ¿quién podría resultar perjudicado si la herramienta se equivoca, si los datos se filtran o si el resultado se malinterpreta?

Los sesgos: de dónde vienen y cómo detectarlos

Los modelos aprenden de textos producidos mayoritariamente en unas pocas lenguas y unas pocas regiones del mundo. De ello resultan **sesgos**:

- **Sesgo lingüístico**: menor calidad en lenguas con poca presencia en Internet (numerosas lenguas africanas, asiáticas o indígenas), y a veces en ciertas variantes del árabe o del francés.
- **Sesgo de perspectiva**: lectura de las crisis «desde el Norte», soluciones estandarizadas, desconocimiento de los actores locales y de las ONG nacionales.
- **Estereotipos**: representaciones de género o culturales (el «médico» es un hombre, la «beneficiaria» es pasiva) en los textos y sobre todo en las imágenes generadas.
- **Sesgo de complacencia**: tendencia a dar la razón a quien pregunta. Si usted pide «confirma que nuestro proyecto es un éxito», el asistente probablemente lo confirmará.

Para detectarlos: pida la posición contraria («¿qué argumentos hay contra esta conclusión?»), haga revisar los textos destinados a las comunidades por personas de esas comunidades y compare con fuentes locales.

Decisiones sobre las personas: el límite que no se debe cruzar

Las decisiones que afectan directamente a los derechos o al acceso a la ayuda de una persona no deben confiarse a un asistente de IA generativa:

- selección o exclusión de beneficiarios;
- evaluación de la credibilidad de un relato (solicitud de asilo, denuncia, caso de protección);
- preselección o rechazo de candidatos a un empleo;
- decisión disciplinaria o de investigación.

Estas herramientas pueden ayudar a **preparar** (redactar una guía de entrevista, estructurar criterios publicados, resumir una política), pero la decisión corresponde a personas habilitadas, según un procedimiento documentado, con posibilidad de recurso. En la Unión Europea, el Reglamento de Inteligencia Artificial (AI Act, adoptado en 2024 y aplicado progresivamente) clasifica además como de «alto riesgo» ciertos usos como la contratación o el acceso a servicios esenciales, con obligaciones reforzadas.

> Prueba: si la persona afectada supiera que la decisión la tomó una IA, ¿sería aceptable y podría impugnarla? Si la respuesta es no, no utilice la IA para decidir.

Transparencia y responsabilidad

La confianza es el capital de una organización humanitaria. Algunas reglas sencillas la protegen:

- **Responsabilidad humana**: cada documento tiene un autor humano responsable de su contenido, haya sido ayudado o no por la IA.
- **Transparencia proporcionada**: indique el uso de la IA cuando sea significativo o cuando el destinatario pueda razonablemente esperarlo (evaluación, investigación, contenido publicado, intercambios con las comunidades). Para una corrección ortográfica, no es necesario.
- **Respeto de las reglas de los donantes y socios**: algunos donantes y convocatorias de propuestas regulan o piden declarar el uso de la IA en las candidaturas; lea las condiciones (semana 6).
- **Sin suplantación**: no genere nunca una imagen, una voz o una cita falsamente atribuida a una persona real, y no presente nunca una imagen generada como una foto de terreno.

Las directrices éticas de comunicación del sector (dignidad, consentimiento informado, sin miserabilismo) se aplican plenamente a los contenidos producidos con la IA.

Derechos de autor, plagio e integridad

Tres puntos de atención:

- **Contenidos ajenos**: pedir a la IA que «reescriba» el informe de otra organización para presentarlo como propio sigue siendo plagio. Cite sus fuentes.
- **Contenidos generados**: el estatus jurídico de los textos e imágenes generados varía según los países y evoluciona. Las condiciones de uso de Claude y de Gemini precisan los derechos del usuario sobre los resultados; para un uso comercial o una publicación importante, consúltelas.
- **Integridad académica y profesional**: en los exámenes de este Mini MBA, como en una prueba de selección, utilizar la IA para responder en su lugar es un fraude. El objetivo es que usted domine las competencias, no solo la herramienta.

Para las fotos, la regla de la organización se mantiene: ninguna foto ajena sin derechos, ninguna foto de un niño identificable sin consentimiento, y una imagen generada se señala claramente como tal.

Coste ambiental y equidad de acceso

La IA consume electricidad y agua en los centros de datos. Las estimaciones por petición varían mucho según los estudios y los modelos; el impacto global del sector crece con fuerza. Para una organización comprometida con el clima, reflejos razonables:

- utilizar la IA cuando aporta un valor real, no por reflejo para tareas triviales;
- preferir un documento bien preparado a decenas de intentos sucesivos;
- integrar el uso de lo digital en la política ambiental.

Equidad de acceso: los equipos de terreno tienen a menudo una conexión débil, dispositivos antiguos y menos acceso a suscripciones de pago. Una organización que reserva la IA a la sede y al personal internacional agranda las desigualdades. Una política justa prevé el acceso, la formación y las lenguas para el personal nacional y los socios locales.

Ejercicio: analizar tres usos desde el punto de vista ético

Situación. Una ONG estudia tres usos: A) generar imágenes «realistas» de niños desnutridos para una campaña de recaudación, a falta de fotos autorizadas; B) pedir a la IA que clasifique 600 candidaturas a un puesto de conductor y rechace automáticamente las 500 peor puntuadas; C) utilizar la IA para proponer una primera versión de los mensajes de sensibilización sobre el cólera en francés fácil, luego probados con grupos de la comunidad y traducidos por hablantes nativos.

Trabajo solicitado. Para cada uso, diga si es aceptable y por qué, y proponga una alternativa si es necesario.

Corrección. A) No aceptable: una imagen generada presentada como real engaña al público, atenta contra la dignidad y refuerza el miserabilismo; alternativa: fotos con consentimiento informado, ilustraciones claramente señaladas o mensajes centrados en las soluciones. B) No aceptable: decisión sobre personas sin control humano, riesgo de sesgo y de discriminación, ausencia de recurso; alternativa: criterios publicados aplicados por un panel, pudiendo la IA ayudar a redactar la rejilla o una prueba práctica. C) Aceptable: la IA prepara un primer borrador, la comunidad y los hablantes nativos lo validan; es un uso proporcionado, transparente y verificado.

Para recordar

- Los principios humanitarios y el principio de no hacer daño se aplican al uso de la IA.
- Los sesgos afectan a la lengua, la perspectiva, los estereotipos y la complacencia; busque la contradicción.
- Ninguna decisión sobre los derechos o el acceso a la ayuda de una persona se confía a una IA generativa.
- Un autor humano sigue siendo responsable; indique el uso significativo de la IA; nunca una imagen generada presentada como real.
- Piense en el coste ambiental y en la equidad de acceso para el personal nacional y los socios.

Cuestionario

1. Pedir a la IA «confirma que nuestro proyecto es un éxito» expone sobre todo al sesgo:
   A. Lingüístico
   B. De complacencia
   C. De fecha
   D. De formato

2. ¿Cuál de estas tareas NO debe confiarse a una IA generativa?
   A. Redactar una guía de entrevista
   B. Resumir una política de RR. HH. publicada
   C. Decidir qué hogares quedan excluidos de la ayuda
   D. Corregir la ortografía de un informe

3. Una imagen generada de una situación de terreno puede utilizarse:
   A. Como foto real si es realista
   B. Nunca, en ningún caso
   C. Solo para los donantes
   D. Si se señala claramente como generada y respeta la dignidad

4. El principio de independencia invita a:
   A. Conservar la capacidad de trabajar sin la herramienta para las funciones críticas
   B. Utilizar un solo proveedor
   C. No utilizar nunca la IA
   D. Dejar que la IA decida

5. Responder a los exámenes de este Mini MBA con la IA en su lugar es:
   A. Recomendable
   B. Un fraude
   C. Obligatorio
   D. Indiferente

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