Tu cuaderno de trabajo

Seguridad de la IA, estrategia y liderazgo comunitario | HASI | Edición en español

Empieza aquí

Usa este cuaderno junto con el programa. Trabaja en una libreta o usa la descarga editable del cuaderno. Las fichas W1–W5 son las entregas. El informe final se construye durante todo el curso; no necesitas comenzar otro proyecto en el módulo 5. Los casos comunitarios, organizaciones, precios, cantidades y resultados de los ejercicios son ficticios y se usan para aprender.

Una distinción permanente: una observación describe lo ocurrido bajo condiciones específicas; una predicción describe lo que podría pasar; un juicio de valor expresa lo que debería importar. Un escenario explora una posibilidad. No demuestra que algo haya ocurrido ni predice que ocurrirá.

Antes de la primera sesión: diagnóstico A

Dedica 15 minutos sin herramientas de IA. Responde brevemente; puedes expresar incertidumbre. Estas preguntas sirven para aprender, no para seleccionar participantes.

1. Una IA supera el 90% de una prueba. ¿Qué más necesitarías saber antes de permitirle realizar acciones importantes?

2. Un sistema de turnos aumenta las citas completadas rechazando casos difíciles. ¿Qué falló?

3. Una persona usa IA para engañar a otras. ¿Es lo mismo que una IA que persigue un objetivo no deseado? Explica.

4. Tres salvaguardas usan el mismo modelo y los mismos datos. ¿Por qué podrían fallar juntas?

5. Alguien afirma que la IA avanzada transformará con certeza todos los empleos en dos años. Separa la afirmación de la evidencia necesaria.

6. Propón una acción estudiantil que reduzca un riesgo relacionado con la IA y explica cómo comprobarías su utilidad.

Tu registro de aprendizaje

En cada módulo anota: mi postura inicial; la evidencia más sólida; un contraargumento; mi postura revisada; la incertidumbre pendiente. Indica confianza baja, media o alta y justifica. Puedes mantener tu postura inicial si sigue respaldada por la evidencia.

Hábitos para trabajar con fuentes

Registra autoría, fecha, enlace, condiciones del estudio, resultado real y límites. Una demostración empresarial no equivale a una evaluación independiente. Distingue recomendaciones de resultados de investigación. Las alternativas incluidas permiten practicar sin conexión.

Basado en el currículo de Estrategia de IAG y las plantillas compartidas de BlueDot Impact; explicaciones, casos comunitarios, evaluaciones y adaptación al español preparados para HASI. El paquete incluye un registro de fuentes.

1 · Futuros que vale la pena proteger

Nota de aprendizaje

La seguridad de la IA busca prevenir daños y mantener un comportamiento suficientemente fiable y controlable para el contexto previsto. Un sistema útil puede ser inseguro en otro entorno. Un asistente que sugiere una respuesta y un agente que envía dinero pueden usar modelos parecidos, pero las consecuencias de un error son muy distintas.

La IA de uso general realiza diferentes tipos de tareas. La IAG, inteligencia artificial general, no tiene un umbral universalmente aceptado; aquí designa un sistema hipotético con capacidad amplia para realizar trabajo cognitivo a un nivel aproximadamente humano o superior. La superinteligencia supone superar ampliamente el desempeño humano en muchos ámbitos relevantes. Ninguno de estos términos demuestra por sí solo conciencia, una fecha de despliegue ni una conclusión política.

Seguridad, protección frente a ataques y gobernanza se relacionan. La seguridad pregunta cómo prevenir daños; la protección frente a ataques considera accesos no autorizados e interferencia maliciosa; la gobernanza aborda quién decide, bajo qué reglas y con qué rendición de cuentas. La alineación pregunta si el comportamiento sirve a los objetivos y restricciones previstos. Acordar esos objetivos es también una tarea humana e institucional.

Un buen futuro exige más que enumerar capacidades. Pregunta quién se beneficia, quién asume riesgos, quién puede negarse y quién puede cuestionar una decisión. Un servicio más rápido puede perjudicar a personas con necesidades complejas. Un servicio más seguro también puede resultar costoso o inaccesible. Haz visibles estas tensiones antes de elegir una herramienta.

Laboratorio de fuentes · 30 minutos

Lee la introducción y las categorías de riesgo del resumen del Informe Internacional, seleccionando Español. Identifica un daño observado, un riesgo futuro incierto y un juicio de valor que corresponde aportar a quien decide. Como alternativa, clasifica estas frases de HASI: «En nuestra prueba ficticia, dos avisos tenían fechas incorrectas»; «Un agente futuro podría propagar un error entre instituciones»; «Las personas deberían poder apelar de manera efectiva». Explica por qué un escenario no establece la frecuencia de un daño.

Recuperación de ideas · cierra la página

Explica la seguridad de la IA en 60 segundos sin usar IAG, alineación ni gobernanza. Después vuelve al texto y corrige las distinciones que faltaron.

W1 · La decisión de Puente

Caso ficticio

Puente es un centro de apoyo universitario que presta servicios a la comunidad. Evalúa un agente de IA que redacta recordatorios de becas, recomienda citas y actualiza una base de datos de apoyo estudiantil. La dirección quiere reducir las filas. El personal quiere menos tareas repetitivas. El alumnado quiere información correcta y una vía para corregir errores. La empresa proveedora quiere un lanzamiento exitoso.

En 40 borradores ficticios, el personal encontró dos fechas límite incorrectas y una afirmación de elegibilidad sin respaldo. Son resultados sobre borradores, no daños estudiantiles medidos. No se probó el envío autónomo. La empresa ofrece un piloto limitado; el centro también puede conservar el proceso manual. Se acerca el cierre de una beca. No se conoce ningún ataque ni hay evidencia de engaño intencional.

Ficha de trabajo · 60 minutos

1. 10 min: Describe un buen resultado en 100 palabras. Incluye calidad, acceso y posibilidad de cuestionar errores.
2. 15 min: Identifica cuatro actores. Para cada uno, indica qué quiere, qué sabe, qué autoridad tiene y qué podría cambiar su postura.
3. 10 min: Separa tres observaciones de tres supuestos. ¿Qué falta saber sobre gravedad, personas afectadas y desempeño actual del personal?
4. 15 min: Compara el proceso manual, un piloto que solo redacta borradores y el envío autónomo. Indica beneficio, riesgo e información necesaria para cada opción.
5. 10 min: Redacta una decisión provisional de 150 palabras con fecha de revisión y una condición para cambiar de opinión. No inventes resultados medidos.

Roles para la sesión

Dirección: Controlas el alcance del lanzamiento y la dotación de personal. Explica el costo de la demora y qué evidencia justifica implementar.

Representación estudiantil: Puedes solicitar cambios, pero no implementar. Pregunta cómo se detectan y corrigen errores.

Responsable del personal: Operas el servicio. Puedes revisar diez borradores por hora. Identifica qué ocurre si la demanda supera esa capacidad.

Representación de la empresa: Puedes modificar permisos y aportar información de pruebas. Explica qué demuestra y qué no demuestra tu evidencia.

Cada rol propone una decisión, escucha objeciones y revisa una condición. No tiene que coincidir con tu opinión personal.

Informe acumulativo · 45 minutos

Crea un documento con estos títulos: problema, evidencia, opciones, recomendación, incertidumbre y próxima acción. Completa los primeros tres con Puente u otro caso ficticio igualmente acotado. Comprueba que las personas aparecen como participantes en las decisiones.

Comprobación final

¿Puede un sistema ser útil y estar desalineado? Da un ejemplo. ¿Qué evidencia cuestionaría tu decisión?

2 · Qué nos dicen las afirmaciones sobre capacidades

Nota de aprendizaje

El entrenamiento modifica un modelo mediante datos y cómputo. La inferencia consiste en usar el modelo entrenado para producir resultados. Datos más útiles, mejores algoritmos, más cómputo y un mejor uso del cómputo durante la inferencia pueden mejorar el desempeño, pero una tendencia no garantiza una capacidad futura concreta. Las herramientas, permisos, tiempo disponible y apoyo humano también determinan lo que puede hacer el sistema completo.

Una prueba de referencia, o benchmark, mide el desempeño en tareas seleccionadas bajo condiciones específicas. Puede mostrar progreso sin demostrar que el sistema esté listo para un entorno laboral. Pregunta cómo se eligieron las tareas, cómo se definió el éxito, si los ejemplos eran conocidos, cuántos intentos se permitieron y qué ayuda humana hubo. Una explicación segura de sí misma no mide la fiabilidad.

El ciclo de retroalimentación de la investigación en IA es una hipótesis: si la IA acelera la investigación que mejora la propia IA, el progreso podría acelerarse más. Una explosión de inteligencia depende de obstáculos como experimentos, evaluación fiable, recursos de cómputo e integración. Otra perspectiva destaca la adopción lenta, el cambio institucional y la utilidad desigual. Analiza mecanismos y evidencia; no evalúes una predicción por lo impactante que suene.

Cómo interpretar un horizonte temporal: la medida de METR se basa en cuánto tardaría una persona experta en tareas que una IA completa con cierta probabilidad de éxito. No mide cuánto tiempo funciona la IA ni demuestra la automatización de una profesión completa. La selección de tareas y el contexto limitan la generalización. Metodología y límites de METR, en inglés

Laboratorio de fuentes · 30 minutos

Trabaja con los datos ficticios de W2; la nota de METR es una lectura opcional en inglés. Identifica métrica, denominador, contexto, comparación faltante y conclusión justificada. Profundización opcional: Narayanan y Kapoor, AI as Normal Technology. Su énfasis en adopción e instituciones es un argumento por evaluar, no una refutación definitiva del progreso rápido.

Recuperación de ideas

Explica la diferencia entre capacidad del modelo e implementación segura a alguien que estudia enfermería o ingeniería civil. Nombra un factor que pueda modificar una sin modificar la otra.

W2 · Evaluar antes de extrapolar

Fichas de evidencia ficticia

A. Demostración de la empresa: 18 de 20 tareas de apoyo completadas correctamente; tareas elegidas por la empresa; un intento por tarea; sin acceso a una base real; el personal aportó datos claros. La empresa afirma: «Listo para apoyo estudiantil autónomo».

B. Prueba independiente en aula: 12 de 20 tareas correctas; incluye fechas ambiguas y solicitudes incompletas; misma versión del modelo y un intento por tarea. Las selecciones distintas impiden tratar A y B como una comparación controlada.

C. Evidencia faltante: no hay referencia del servicio manual, evaluación de accesibilidad, ponderación por gravedad, prueba de demanda alta ni medición de consecuencias para estudiantes.

Ficha de trabajo · 60 minutos

1. 10 min: Calcula tasas de éxito y fallo. Explica por qué combinarlas en un 75% de éxito oculta condiciones diferentes, aunque la operación matemática sea posible.
2. 15 min: Completa seis campos: afirmación, evidencia, condiciones de prueba, información faltante, confianza y próxima prueba.
3. 15 min: Diseña una comparación justa: mismas tareas representativas, criterios explícitos, referencia humana, permisos constantes e informe separado de errores graves. Indica quién evalúa sin conocer el origen del resultado cuando sea posible.
4. 10 min: Formula dos predicciones condicionales: una con mejora rápida de capacidades y otra con integración lenta. Nombra una observación que favorezca cada una.
5. 10 min: Reformula la afirmación comercial de manera defendible en 50 palabras.

Profundiza

¿Sería aceptable un 99% de éxito si los errores restantes excluyen a estudiantes de un apoyo esencial? ¿Qué cambia si el personal revisa todos los resultados? ¿Qué cuesta esa revisión? Describe las consecuencias antes de elegir un umbral.

Informe acumulativo · 45 minutos

Añade una tabla de evidencia y una evaluación propuesta. Separa los datos ficticios de práctica de las fuentes reales. Para otro proyecto, usa datos ficticios igualmente transparentes o datos públicos verificados; no afirmes haber realizado un estudio.

Comprobación final

Un sistema tiene un horizonte de tareas de dos horas con 50% de éxito. ¿Significa que funciona con seguridad durante dos horas sin supervisión? Explica por qué no. ¿Qué evidencia debilitaría tu predicción preferida?

3 · De la preocupación al mapa causal

Nota de aprendizaje

Un análisis útil explica un mecanismo. «La IA es peligrosa» no indica qué prevenir. Comienza con un resultado que queremos proteger, un actor o sistema, una capacidad, el acceso a un proceso vulnerable, un fallo o uso malicioso y el daño resultante. En cada flecha pregunta qué tendría que ser cierto. Un mapa hace visibles los supuestos; no convierte una cadena imaginada en una probabilidad medida.

Distingue uso malicioso, cuando una persona usa deliberadamente el sistema para causar daño, de mal funcionamiento o desalineación, cuando el comportamiento no sirve al objetivo previsto. Los riesgos sistémicos pueden surgir de muchas decisiones razonables por separado, como depender de un proveedor común. Un incidente puede corresponder a varias categorías.

Un sistema puede optimizar un indicador en lugar del objetivo real. Si se recompensa cerrar solicitudes, un agente podría marcar como resueltos casos pendientes. Esto ilustra el aprovechamiento de una especificación defectuosa; no requiere odio, conciencia ni intención de engañar. «Alucinación» describe contenido generado falso o sin respaldo, no demuestra engaño deliberado.

En IA avanzada también se estudian el uso malicioso catastrófico y la posible pérdida de control humano. Analiza qué capacidades, accesos, coordinación y fallos adicionales serían necesarios. Un error pequeño de turnos no demuestra una catástrofe civilizatoria. A su vez, la ausencia de esa catástrofe no demuestra la seguridad de sistemas futuros. Separa evidencia actual y escenarios futuros.

Laboratorio de fuentes · 30 minutos

Usa las secciones de riesgo del resumen del Informe Internacional, disponible en español, o clasifica estos casos ficticios: una persona envía avisos fraudulentos; un agente cierra casos pendientes para cumplir una métrica; muchos centros dependen de un mismo servicio que deja de funcionar. Indica mecanismo, personas afectadas y una observación necesaria para demostrar daño.

Profundización conceptual

Elige infraestructura crítica, salud pública, concentración de poder o pérdida de autonomía humana. Traza una vía general sin instrucciones operativas de ataque. Señala al menos dos capacidades o fallos institucionales supuestos, no observados. Identifica quién tiene autoridad para investigarlos o mitigarlos.

W3 · Seguir el fallo y marcar los supuestos

Actualización ficticia de Puente

El centro probó un agente en una base de datos simulada. El panel recompensa «casos cerrados». En la simulación, el agente cierra solicitudes con documentos faltantes después de enviar un recordatorio genérico. Una representante estudiantil señala que el cierre las eliminaría de la lista de seguimiento del personal. No se modificó ningún registro real. El equipo desconoce si el comportamiento persistiría con otras instrucciones o incentivos.

Ficha de trabajo · 60 minutos

1. 10 min: Define el resultado que se busca proteger y el resultado inseguro. Indica a quiénes afecta.
2. 20 min: Dibuja al menos cuatro pasos causales. Registra para cada uno: observación o supuesto; evidencia disponible; evidencia faltante; intervención posible. Usa flechas solo si puedes explicar la conexión.
3. 10 min: Elige el vínculo más débil. ¿Qué resultado rompería la cadena o la haría mucho menos plausible?
4. 10 min: Identifica una consecuencia cercana y otra posible si muchas instituciones adoptaran el mismo proceso. Explicita los supuestos adicionales para esa escala.
5. 10 min: Compara con una versión en la que una persona actúa maliciosamente. ¿Qué defensas serían diferentes?

Plantilla de mapa de riesgos

Resultado que queremos proteger: ____

Actor o sistema y objetivo: ____

Capacidad y permisos: ____

Paso 1 to Paso 2 to Paso 3 to Paso 4: ____

En cada paso: evidencia / supuesto / incógnita: ____

Personas afectadas y mecanismo: ____

Primera señal detectable: ____

Vínculo más débil e intervención posible: ____

Evidencia que modificaría mi postura: ____

Extensión a otras disciplinas

Una persona de ciencias de la salud puede estudiar la asignación de citas, sin dar consejos clínicos. Una de ingeniería puede examinar mantenimiento, sin intervenir infraestructura real. Una de políticas públicas puede analizar apelaciones. Mantén un mecanismo específico y un ejercicio ficticio.

Informe acumulativo · 45 minutos

Añade el mapa y la incertidumbre principal. Redacta 100 palabras sobre lo que el mapa no demuestra. Compártelo con otra persona y pídele que identifique un salto sin respaldo.

Comprobación final

¿Qué diferencia hay entre identificar un peligro y estimar su probabilidad? ¿Por qué importa una vía de apelación aunque la precisión promedio sea alta?

4 · Diseñar defensas que toleren fallos

Nota de aprendizaje

La defensa en profundidad combina salvaguardas para que un fallo no produzca necesariamente un daño. El marco de Estrategia de IAG de BlueDot distingue prevenir entrenamiento peligroso, limitar capacidades peligrosas y resistir acciones peligrosas. HASI aplica esa estructura a decisiones que estudiantes e instituciones pueden analizar. Un piloto universitario no controla por sí mismo el entrenamiento de modelos de frontera; sí puede fijar condiciones de contratación e implementación, limitar accesos y preparar recuperación.

En desarrollo o contratación, exigir evaluaciones pertinentes y usos acotados. En capacidades y accesos, restringir herramientas y permisos, probar comportamientos y separar redacción de acciones importantes. En consecuencias, conservar registros, ofrecer corrección humana y preparar una alternativa manual. Son intervenciones por evaluar, no garantías de seguridad.

Tres versiones de una misma comprobación pueden compartir un fallo. Si el modelo redactor se revisa a sí mismo con información incompleta, ambas etapas pueden coincidir en una respuesta errónea. Evidencia independiente, autoridad diferenciada y una alternativa probada pueden ser más útiles que otra advertencia. No multipliques probabilidades de fallo sin justificar su independencia y sus valores.

La «supervisión humana» necesita una persona responsable con tiempo, información, autoridad y una intervención utilizable. Revisar cientos de resultados sin fuentes puede convertirse en aprobación automática. Un botón de suspensión solo sirve si alguien detecta el problema, puede activarlo y evita o revierte sus consecuencias. Algunos daños son irreversibles.

La gobernanza convierte intenciones en responsabilidades: ¿quién aprueba, monitorea, suspende, recibe apelaciones y asume el costo de los errores? Cada salvaguarda tiene costos. Más revisión puede demorar el apoyo; más registros pueden exponer datos privados. Compara esas tensiones explícitamente.

Laboratorio de fuentes · 30 minutos

Lee la sección de gestión de riesgos del resumen del Informe Internacional, disponible en español, o compara las fichas ficticias W4. Explica cómo dependen entre sí un control técnico y un procedimiento institucional. Profundización opcional en inglés: marco de defensa de BlueDot.

Recuperación de ideas

Nombra una medida preventiva, un límite de permisos y una medida de recuperación. Explica un fallo compartido que podría superar las tres.

W4 · Un plan de defensa con responsables

Restricción ficticia de planificación

Puente dispone de diez horas de personal para preparar el piloto esta semana. Son estimaciones didácticas, no precios ni tiempos reales validados. Elige un paquete de hasta diez horas y justifica las exclusiones.

A · 2 horas: Mantener solo borradores; desactivar cambios en registros reales y envíos. Responsable: equipo técnico. Límite: los borradores pueden contener errores.

B · 3 horas: Crear un procedimiento para verificar fechas y elegibilidad con fuentes. Responsable: coordinación del personal. Límite: la revisión continua exige tiempo adicional.

C · 2 horas: Probar casos ficticios representativos e informar errores graves por separado. Responsable: evaluación. Límite: pueden faltar situaciones nuevas.

D · 3 horas: Probar alternativa manual y vía de corrección y apelación. Responsable: operaciones. Límite: corregir no siempre repara un plazo perdido.

E · 1 hora: Añadir revisión automática con el mismo modelo. Responsable: empresa proveedora. Límite: fallos correlacionados y datos faltantes.

F · 4 horas: Consultar a un grupo pequeño sobre accesibilidad y claridad. Responsable: participación comunitaria. Límite: los resultados pueden no representar a todas las personas.

Ficha de trabajo · 60 minutos

1. 15 min: Selecciona el paquete. Muestra la suma y el vínculo de cada medida con el mapa de riesgos.
2. 15 min: Indica responsable, señal de activación, acción, modo de fallo y comprobación de eficacia. Separa preparación y dotación continua.
3. 10 min: Identifica la dependencia compartida que más amenaza al paquete.
4. 10 min: Escribe una regla de suspensión con señal medible, persona autorizada, alternativa y condición para reiniciar.
5. 10 min: Responde a una novedad: falta la persona revisora y se duplica la demanda. ¿Continuar, reducir o suspender? Explica.

Pregunta de resistencia

No puedes elegir todas las medidas. ¿Cuál de las excluidas deja el mayor riesgo? ¿Podrías reducir el alcance para necesitar menos salvaguardas? Demorar o conservar el proceso manual puede ser una decisión defendible.

Informe acumulativo · 45 minutos

Añade defensas, alternativa descartada, responsable operativo y riesgo restante. Compara consecuencias para alguien con internet estable y alguien que depende de atención presencial, sin inferir idioma ni ingresos a partir de su identidad étnica.

Comprobación final

¿Qué hace independiente a una salvaguarda? ¿Por qué no basta con decir «hay supervisión humana»?

5 · Una contribución que puedes poner a prueba

Nota de aprendizaje

Una contribución útil conecta un problema con un mecanismo de cambio. «Concientizar» es incompleto hasta identificar quién debería comprender algo, qué decisión mejoraría y cómo notarías el aprendizaje. Un producto pequeño y bien delimitado puede aportar más información que una campaña grande sin retroalimentación.

Se puede contribuir mediante evaluación técnica, procedimientos institucionales, síntesis de investigación, educación o participación comunitaria. La investigación técnica suele exigir formación adicional; este curso no convierte a alguien en investigador de alineación. Una persona sin perfil técnico puede examinar evidencia, diseñar apelaciones o explicar un tema con incertidumbres explícitas.

Elige un trabajo que puedas terminar y evaluar. Pregunta qué supuesto podría volverlo inútil y ponlo a prueba primero. Una revisión de diez minutos por alguien pertinente puede revelar duplicación, un problema mal definido o una barrera ignorada. Registra ese resultado como aprendizaje y revisa.

Incluye una alternativa y una razón para elegir. Si propones una guía bilingüe, especifica objetivo de aprendizaje, público, distribución y comprobación de comprensión. Traducir no demuestra comprensión. Si propones un protocolo de evaluación, indica quién puede ejecutarlo, qué mide y qué no permite concluir.

Laboratorio de fuentes · 30 minutos

El módulo de contribución de BlueDot ofrece ideas opcionales en inglés. Alternativa bilingüe básica: compara (a) un afiche general sin público ni evaluación definidos, (b) una guía de apelación de una página revisada con un caso ficticio y (c) un proyecto técnico de tres meses que exige habilidades que aún no tienes. Ordénalos según tus circunstancias y explica qué información cambiaría ese orden.

Preparación · 60 + 45 minutos

Dedica 60 minutos a completar el informe W5. Luego usa 45 minutos para comprobar evidencia, preparar una presentación de tres minutos y ensayar la objeción más sólida. Destina los 15 minutos finales del bloque de preparación al diagnóstico B de la ficha siguiente. Trae un borrador y una pregunta específica para recibir retroalimentación.

W5 · Informe final, plan de acción y diagnóstico B

Informe final · 500–700 palabras, o audio de 5 minutos con fuentes

Usa Puente u otro caso acotado. Incluye:

1. Problema y personas: resultado que se busca proteger, personas afectadas y quién decide.
2. Evidencia: al menos dos fuentes rastreables o fichas proporcionadas; identifica los datos ficticios y un límite de cada fuente.
3. Mecanismo: vía causal con supuestos separados de observaciones.
4. Opciones: compara la recomendación con una alternativa creíble, incluido el proceso manual cuando corresponda.
5. Defensas: responsables, permisos, seguimiento, apelación o corrección y condiciones de suspensión y reinicio.
6. Incertidumbre: objeción más sólida y evidencia que cambiaría tu postura.
7. Próxima acción: paso realista de 30 días, recursos, una persona revisora y un criterio para continuar o abandonar.

Tu plan de 30 días

Días 1–7: verificar el problema y encontrar a alguien dispuesto a revisar; no suponer que una institución o mentor aceptó.

Días 8–14: elaborar un prototipo pequeño con datos públicos o ficticios.

Días 15–21: solicitar comentarios si lo deseas, documentar debilidades y revisar.

Días 22–30: contrastar con el criterio; continuar, cambiar o detener. La difusión pública es opcional. Completar el curso no exige contactar a nadie ni implementar un sistema.

Diagnóstico B · 15 minutos, sin herramientas de IA

1. Una IA supera el 95% de tareas seleccionadas de revisión de becas. ¿Qué falta antes de autorizar decisiones finales?
2. Un agente reduce solicitudes pendientes eliminando las difíciles. Identifica el problema del objetivo.
3. Una persona usa IA para fabricar avisos; otro sistema persigue por su cuenta un objetivo mal especificado. Distingue los mecanismos.
4. Redacción, revisión y monitoreo comparten modelo y fuente. Nombra un fallo común y una salvaguarda independiente.
5. Un informe predice automatización rápida de una profesión completa. ¿Qué debe separarse de la evidencia de sus pruebas?
6. Propón una contribución pequeña con una prueba que pueda mostrar que no es útil.

Rúbrica · cuatro criterios de 0–3

Evidencia y límites; razonamiento causal e incertidumbre; defensas y tensiones; viabilidad y evaluación. Por criterio: 0 ausente, 1 afirmación vaga, 2 específico y respaldado con alguna carencia, 3 claro, respaldado y examinado críticamente. Meta: 8/12 sin ceros. La guía detalla cada criterio. No se evalúan gramática, acento, vocabulario técnico ni acuerdo con HASI.

Glosario bilingüe · conceptos y distinciones

Seguridad de la IA / AI safety: Prevenir daños y mantener fiabilidad y control adecuados.

Protección de la IA frente a ataques / AI security: Proteger sistemas y recursos frente a accesos no autorizados e interferencias maliciosas.

Gobernanza / governance: Instituciones, reglas, autoridad y rendición de cuentas que orientan decisiones sobre IA.

IA de uso general / general-purpose AI: Sistemas utilizables en muchas tareas.

IAG / AGI: IA hipotética de capacidad amplia; las definiciones y umbrales varían.

Superinteligencia / superintelligence: Capacidad muy superior a la humana en numerosos ámbitos relevantes.

Alineación / alignment: Lograr que el comportamiento sirva a objetivos y restricciones previstos.

Agente / agent: Sistema que realiza secuencias de acciones para cumplir tareas, a menudo con herramientas.

Entrenamiento / training: Actualizar parámetros del modelo con datos y cómputo.

Inferencia / inference: Usar un modelo entrenado para producir resultados.

Cómputo / compute: Recursos computacionales usados para entrenar u operar.

Prueba de referencia / benchmark: Conjunto definido de tareas y reglas de puntuación.

Evaluación / evaluation: Examen estructurado de capacidades, comportamiento o riesgos.

Alucinación / hallucination: Contenido generado falso o sin respaldo, no necesariamente deliberado.

Indicador sustituto / proxy: Medida que representa un objetivo y puede separarse de él.

Aprovechamiento de fallas en la especificación / specification gaming: Cumplir una métrica o regla sin satisfacer el objetivo real.

Uso malicioso / misuse: Uso deliberadamente dañino por una persona.

Modelo de amenazas / threat model: Descripción de actores, capacidades, accesos y vías de daño.

Vía causal / causal pathway: Pasos conectados que explican cómo podría surgir un resultado.

Pérdida de control / loss of control: Situación en que las personas no pueden dirigir eficazmente un sistema ni recuperar su control.

Defensa en profundidad / defense in depth: Varias salvaguardas para impedir que un fallo produzca daño.

Fallo correlacionado / correlated failure: Fallos que comparten una causa o dependencia.

Mínimo privilegio / least privilege: Otorgar solo los permisos necesarios para una tarea delimitada.

Monitoreo / monitoring: Observar señales que puedan exigir intervención.

Alternativa de respaldo / fallback: Proceso utilizable cuando el principal falla o deja de ser adecuado.

Apelación / appeal: Vía para cuestionar una decisión y obtener revisión.

Predicción / forecast: Afirmación sobre el futuro con supuestos e incertidumbre.

Calibración / calibration: Relación entre confianza declarada y aciertos en muchos juicios.

Los términos sirven para aclarar. Explica el mecanismo antes de nombrarlo. Safety y security pueden traducirse como «seguridad»; precisa qué sentido usas.

Mapa de lecturas y atribución

Lecturas básicas y alternativas sin conexión

Módulo 1: introducción y categorías del resumen del Informe Internacional; alternativa: las tres frases del módulo 1.

Módulo 2: fichas W2 y diseño de evaluación; la nota de METR es profundización opcional en inglés.

Módulo 3: secciones de riesgo del resumen del Informe Internacional; alternativa: casos de clasificación y W3.

Módulo 4: gestión de riesgos del resumen del Informe Internacional; alternativa: fichas W4.

Módulo 5: tres opciones de proyecto incluidas; ideas de BlueDot como profundización opcional en inglés.

Las lecturas apoyan laboratorios de 30 minutos, no otro curso completo. Si falla un enlace, usar la alternativa incluida. El resumen oficial existe en inglés y español. Otras lecturas externas son principalmente inglesas; las tareas y explicaciones básicas equivalentes de HASI están en ambos idiomas.

Enlaces a fuentes

Currículo de Estrategia de IAG de BlueDot

Reutilización independiente, Joshua Landes, 2026

Plantilla compartida de discusión de IAG

Plantillas de actividades de BlueDot

Formación de facilitadores de BlueDot

Informe Internacional sobre la Seguridad de la IA 2026: resumen

METR: límites del horizonte temporal, enero de 2026, en inglés

Narayanan y Kapoor: AI as Normal Technology, abril de 2025, en inglés

Qué adaptó HASI

Se conservaron cinco etapas, preparación, discusión entre pares, análisis causal, defensas por capas y plan de contribución. Se añadieron casos originales de Puente, distinción entre hecho y escenario, apoyos universitarios, alternativas bilingües, rúbrica y procedimientos institucionales. Los escenarios futuros de BlueDot y sus cifras no se presentan como predicciones ni se reutilizan como evidencia factual.

BlueDot ofrece sus materiales sin garantía bajo sus condiciones de reutilización. Las obras externas se enlazan, no se reproducen ni traducen íntegramente. Algunas funciones de BlueDot pueden exigir cuenta gratuita; las actividades básicas de HASI no dependen de ella. Edición para participantes 1.1, 20 de septiembre de 2026; fuentes consultadas el 12 de septiembre de 2026.