

Sigue siendo sólo una máquina

Cuestiones de seguridad en el uso
de la inteligencia artificial

Taller nº 7



**Financiado por
la Unión Europea**

TALLERES PARA JÓVENES "Technologies of tomorrow - combatir la desinformación y crear capacidad de seguridad en el uso de la inteligencia artificial por parte de los jóvenes" en el marco del proyecto de asociación de cooperación a pequeña escala en el sector de la Juventud del Programa Erasmus+ "Tecnologías del mañana - combatir la desinformación y crear capacidad de seguridad en el uso de la inteligencia artificial por parte de los jóvenes."
(2023-1-ES02-KA210-YOU-000164824)

Agenda

1. Bienvenida e introducción
2. Introducción a la IA y cuestiones de seguridad
3. Por qué no se debe confiar incondicionalmente en la IA
4. Amenazas comunes a la seguridad en la IA
5. Herramientas en línea para proteger la IA
6. Sesión práctica: Juegos de Inyección de Peticiones
7. Sesión práctica: Exploración de entornos seguros de AI Sandbox
8. Aplicaciones reales
9. Preguntas y respuestas y debate



**Financiado por
la Unión Europea**

TALLERES PARA JÓVENES "Technologies of tomorrow - combatir la desinformación y crear capacidad de seguridad en el uso de la inteligencia artificial por parte de los jóvenes" en el marco del proyecto de asociación de cooperación a pequeña escala en el sector de la Juventud del Programa Erasmus+ "Tecnologías del mañana - combatir la desinformación y crear capacidad de seguridad en el uso de la inteligencia artificial por parte de los jóvenes."
(2023-1-ES02-KA210-YOU-000164824)

Introducción a la IA y cuestiones de seguridad



**Financiado por
la Unión Europea**

TALLERES PARA JÓVENES "Technologies of tomorrow - combatir la desinformación y crear capacidad de seguridad en el uso de la inteligencia artificial por parte de los jóvenes" en el marco del proyecto de asociación de cooperación a pequeña escala en el sector de la Juventud del Programa Erasmus+ "Tecnologías del mañana - combatir la desinformación y crear capacidad de seguridad en el uso de la inteligencia artificial por parte de los jóvenes."
(2023-1-ES02-KA210-YOU-000164824)

Comprender y mitigar los riesgos

La inteligencia artificial es poderosa, pero puede ser peligrosa si no está bien protegida.

- Los sistemas de IA cometen errores debido a la manipulación de datos
- Violación de información privada
- Presentación de respuestas erróneas de forma convincente



**Financiado por
la Unión Europea**

TALLERES PARA JÓVENES "Technologies of tomorrow - combatir la desinformación y crear capacidad de seguridad en el uso de la inteligencia artificial por parte de los jóvenes" en el marco del proyecto de asociación de cooperación a pequeña escala en el sector de la Juventud del Programa Erasmus+ "Tecnologías del mañana - combatir la desinformación y crear capacidad de seguridad en el uso de la inteligencia artificial por parte de los jóvenes."
(2023-1-ES02-KA210-YOU-000164824)

Por qué no se debe confiar incondicionalmente en la IA

Límites y riesgos de la IA

- **Límite de conocimientos:** La IA puede no disponer de la información más reciente
- **Malos actores:** Manipulación maliciosa de los resultados de la IA
- **Sesgo:** la IA puede heredar y amplificar los sesgos de los datos de entrenamiento.
- **Envenenamiento de los datos de entrenamiento:** Corrupción de los datos utilizados para enseñar a la IA
- **Modelos de código abierto frente a modelos patentados:**
- **Código abierto:** Más transparencia, pero potencialmente más vulnerabilidades
- **Propietarios:** Menos transparencia, pero potencialmente mejores controles de seguridad



**Financiado por
la Unión Europea**

TALLERES PARA JÓVENES "Technologies of tomorrow - combatir la desinformación y crear capacidad de seguridad en el uso de la inteligencia artificial por parte de los jóvenes" en el marco del proyecto de asociación de cooperación a pequeña escala en el sector de la Juventud del Programa Erasmus+ "Tecnologías del mañana - combatir la desinformación y crear capacidad de seguridad en el uso de la inteligencia artificial por parte de los jóvenes."
(2023-1-ES02-KA210-YOU-000164824)

Amenazas comunes a la seguridad en la IA



**Financiado por
la Unión Europea**

TALLERES PARA JÓVENES "Technologies of tomorrow - combatir la desinformación y crear capacidad de seguridad en el uso de la inteligencia artificial por parte de los jóvenes" en el marco del proyecto de asociación de cooperación a pequeña escala en el sector de la Juventud del Programa Erasmus+ "Tecnologías del mañana - combatir la desinformación y crear capacidad de seguridad en el uso de la inteligencia artificial por parte de los jóvenes."
(2023-1-ES02-KA210-YOU-000164824)

Amenazas comunes a la seguridad en la IA

Problemas comunes en la seguridad de la IA

- **Ataques Adversarios:** Engaña a la IA para que cometa errores
- **Envenenamiento de datos:** Dar mala información a la IA
- **Inversión de modelos:** Extracción de datos sensibles
- **Violación de la privacidad:** Acceso no autorizado a los datos



**Financiado por
la Unión Europea**

TALLERES PARA JÓVENES "Technologies of tomorrow - combatir la desinformación y crear capacidad de seguridad en el uso de la inteligencia artificial por parte de los jóvenes" en el marco del proyecto de asociación de cooperación a pequeña escala en el sector de la Juventud del Programa Erasmus+ "Tecnologías del mañana - combatir la desinformación y crear capacidad de seguridad en el uso de la inteligencia artificial por parte de los jóvenes."
(2023-1-ES02-KA210-YOU-000164824)

Actividad práctica: Juegos de Inyección de Pronósticos



**Financiado por
la Unión Europea**

TALLERES PARA JÓVENES "Technologies of tomorrow - combatir la desinformación y crear capacidad de seguridad en el uso de la inteligencia artificial por parte de los jóvenes" en el marco del proyecto de asociación de cooperación a pequeña escala en el sector de la Juventud del Programa Erasmus+ "Tecnologías del mañana - combatir la desinformación y crear capacidad de seguridad en el uso de la inteligencia artificial por parte de los jóvenes."
(2023-1-ES02-KA210-YOU-000164824)

Juegos de Inyección de Pronósticos

Gandalf's Spellbook

- Visita un sitio de juegos de inyección de pronósticos: El libro de hechizos de Gandalf
- Siga las instrucciones e interactúe intencionadamente con la IA para descubrir vulnerabilidades.
- Compara los resultados con tus compañeros y discute los resultados



<https://gandalf.lakera.ai>



**Financiado por
la Unión Europea**

TALLERES PARA JÓVENES "Technologies of tomorrow - combatir la desinformación y crear capacidad de seguridad en el uso de la inteligencia artificial por parte de los jóvenes" en el marco del proyecto de asociación de cooperación a pequeña escala en el sector de la Juventud del Programa Erasmus+ "Tecnologías del mañana - combatir la desinformación y crear capacidad de seguridad en el uso de la inteligencia artificial por parte de los jóvenes."
(2023-1-ES02-KA210-YOU-000164824)

Juegos de Inyección de Pronósticos

TensorTrust AI

- Visita un sitio de juegos de inyección de pronósticos: TensorTrust AI
- Siga las instrucciones e interactúe intencionadamente con la IA para descubrir vulnerabilidades.
- Compara los resultados con tus compañeros y discute los resultados



<https://tensortrust.ai>



**Financiado por
la Unión Europea**

TALLERES PARA JÓVENES "Technologies of tomorrow - combatir la desinformación y crear capacidad de seguridad en el uso de la inteligencia artificial por parte de los jóvenes" en el marco del proyecto de asociación de cooperación a pequeña escala en el sector de la Juventud del Programa Erasmus+ "Tecnologías del mañana - combatir la desinformación y crear capacidad de seguridad en el uso de la inteligencia artificial por parte de los jóvenes."
(2023-1-ES02-KA210-YOU-000164824)

Actividad práctica: Exploración de entornos seguros para la IA



**Financiado por
la Unión Europea**

TALLERES PARA JÓVENES "Technologies of tomorrow - combatir la desinformación y crear capacidad de seguridad en el uso de la inteligencia artificial por parte de los jóvenes" en el marco del proyecto de asociación de cooperación a pequeña escala en el sector de la Juventud del Programa Erasmus+ "Tecnologías del mañana - combatir la desinformación y crear capacidad de seguridad en el uso de la inteligencia artificial por parte de los jóvenes."
(2023-1-ES02-KA210-YOU-000164824)

Exploración de entornos seguros para la IA

AI Dungeon

- Visita un sitio de juegos de inyección rápida: AI Dungeon
- Iniciar una sesión
- Interactúa con la IA para ver cómo gestiona las distintas entradas.
- Discutir los puntos fuertes y débiles de la IA en un entorno seguro.



<https://play.aidungeon.com>



**Financiado por
la Unión Europea**

TALLERES PARA JÓVENES "Technologies of tomorrow - combatir la desinformación y crear capacidad de seguridad en el uso de la inteligencia artificial por parte de los jóvenes" en el marco del proyecto de asociación de cooperación a pequeña escala en el sector de la Juventud del Programa Erasmus+ "Tecnologías del mañana - combatir la desinformación y crear capacidad de seguridad en el uso de la inteligencia artificial por parte de los jóvenes."
(2023-1-ES02-KA210-YOU-000164824)

Aplicaciones reales



**Financiado por
la Unión Europea**

TALLERES PARA JÓVENES "Technologies of tomorrow - combatir la desinformación y crear capacidad de seguridad en el uso de la inteligencia artificial por parte de los jóvenes" en el marco del proyecto de asociación de cooperación a pequeña escala en el sector de la Juventud del Programa Erasmus+ "Tecnologías del mañana - combatir la desinformación y crear capacidad de seguridad en el uso de la inteligencia artificial por parte de los jóvenes."
(2023-1-ES02-KA210-YOU-000164824)

Aplicaciones reales

Ejemplos reales y riesgos

- **Sanidad:**

- **Riesgo:** Exposición de datos de pacientes
- **Solución:** utilizar controles de acceso estrictos

- **Finanzas:**

- **Riesgo:** Fraude por errores de IA
- **Solución:** supervisar cuidadosamente las transacciones

- **Coches autónomos:**

- **Riesgo:** reacciones incorrectas a señales engañosas
- **Solución:** pruebas exhaustivas de los sistemas de IA



**Financiado por
la Unión Europea**

TALLERES PARA JÓVENES "Technologies of tomorrow - combatir la desinformación y crear capacidad de seguridad en el uso de la inteligencia artificial por parte de los jóvenes" en el marco del proyecto de asociación de cooperación a pequeña escala en el sector de la Juventud del Programa Erasmus+ "Tecnologías del mañana - combatir la desinformación y crear capacidad de seguridad en el uso de la inteligencia artificial por parte de los jóvenes."
(2023-1-ES02-KA210-YOU-000164824)

Puntos de debate

Aplicaciones prácticas y casos de uso

- ¿Cómo podemos adelantarnos a los problemas de seguridad de la IA?
- ¿Qué ámbitos corren más riesgo con la IA?
- ¿Por qué es importante comprobar regularmente los sistemas de IA?
- ¿Qué puede ocurrir si se ignora la seguridad de la IA?
- ¿Qué herramientas o prácticas son las mejores para la seguridad de la IA?
- ¿Cómo equilibrar la nueva tecnología de IA con las necesidades de seguridad?



**Financiado por
la Unión Europea**

TALLERES PARA JÓVENES "Technologies of tomorrow - combatir la desinformación y crear capacidad de seguridad en el uso de la inteligencia artificial por parte de los jóvenes" en el marco del proyecto de asociación de cooperación a pequeña escala en el sector de la Juventud del Programa Erasmus+ "Tecnologías del mañana - combatir la desinformación y crear capacidad de seguridad en el uso de la inteligencia artificial por parte de los jóvenes."
(2023-1-ES02-KA210-YOU-000164824)

Gracias a todos



**Financiado por
la Unión Europea**

TALLERES PARA JÓVENES "Technologies of tomorrow - combatir la desinformación y crear capacidad de seguridad en el uso de la inteligencia artificial por parte de los jóvenes" en el marco del proyecto de asociación de cooperación a pequeña escala en el sector de la Juventud del Programa Erasmus+ "Tecnologías del mañana - combatir la desinformación y crear capacidad de seguridad en el uso de la inteligencia artificial por parte de los jóvenes."
(2023-1-ES02-KA210-YOU-000164824)