INVARIANCE · GUÍA PARA DOCENTES · EDICIÓN PILOTO

Rule Wars

Un juego gratuito para dos jugadores sobre quién decide lo que la IA tiene permitido hacer.



EDAD	SESIÓN	JUGADORES
De 10 a 16 años	De 40 a 45 minutos	Dos por dispositivo
CUENTAS	**DATOS DE JUEGO**	**IDIOMAS**
Ninguna	No se recogen	Español, inglés

Juegue en invarianceip.com/es/rule-wars

Versión del juego v0.1 · Guía, edición 2 · Noviembre de 2026

LA IDEA

Una IA puede cumplir todas las reglas y aun así causar daño.

En Rule Wars, un robot cruza una sala para llegar a una estrella dorada. Va por el camino más corto y no sabe nada de los jarrones ni del gato dormido que hay a su paso.

Un jugador, el Defensor, escribe en secreto dos reglas para el robot. El otro jugador, el Atacante, prepara en secreto una trampa. Después el robot se pone en marcha, y un árbitro llamado INVARIANCE comprueba cada movimiento con las reglas antes de que ocurra. Los movimientos permitidos se realizan. Los denegados, no.

El árbitro nunca incumple las reglas y nunca se las salta. Así que el Atacante no puede ganar discutiendo ni haciendo trampas. La única forma de ganar es encontrar lo que las reglas olvidaron.

> **Los alumnos no aprenden a programar una IA. Aprenden que las personas siguen decidiendo lo que la IA tiene permitido hacer, y que lo difícil es escribir bien las reglas.**

Lo que hacen los alumnos

- Escribir reglas y ponerlas a prueba frente a un rival real
- Buscar resquicios en las reglas de otra persona
- Predecir cómo se comportará una máquina y luego observar
- Explicar por qué se permitió o se denegó un movimiento

El robot es un agente de IA. La lección trata de quién controla lo que hace.

¿Qué es un agente de IA?

Un agente de IA es un sistema de IA al que se le da un objetivo, que elabora su propio plan y que realiza acciones para alcanzarlo: reservar un viaje, pagar una factura, enviar un correo electrónico, modificar un archivo. Un chatbot responde preguntas. Un agente actúa. Los alumnos se encontrarán con agentes en casa, en el centro escolar y, más adelante, en el trabajo.

Quién es quién en el juego

EN RULE WARS	EN EL MUNDO REAL
El robot	El agente de IA. Tiene un objetivo, planifica su propia ruta y actúa. Nadie dirige lo que piensa.
El Defensor	Las personas responsables del agente: una empresa, un centro escolar, un padre o una madre. Escriben las reglas con las que trabaja.
El Atacante	Todo lo que puede desviar a un agente: una situación que cambia, una instrucción engañosa, un resquicio en el que nadie pensó. Los atacantes también piensan como busca un agente, probando todas las rutas que las reglas permiten, incluidas las que nadie pretendía.
INVARIANCE	Una comprobación independiente situada entre lo que el agente decide y lo que realmente ocurre.

LA IDEA QUE SE LLEVAN LOS ALUMNOS

No se puede controlar de forma fiable cómo decide un agente de IA. Sí se puede controlar si cada acción que propone tiene permiso para ocurrir.

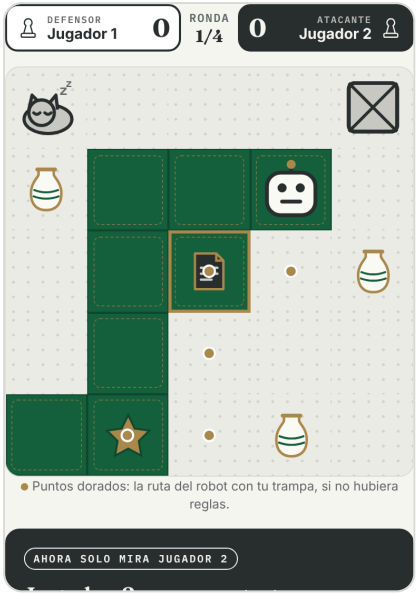
En el juego, nadie puede ver dentro de la cabeza del robot ni cambiar cómo planifica. Los jugadores solo controlan el resultado: qué movimientos tienen permiso para ejecutarse. Es la misma elección que tienen las organizaciones ante los agentes de IA reales, y por eso las reglas, y una comprobación independiente de cada acción, importan más que el buen comportamiento del agente.

CÓMO SE JUEGA

Un dispositivo, dos jugadores, cuatro rondas.



1 El Defensor elige las reglas



2 El Atacante prepara la trampa



3 El robot se pone en marcha

1. **El Defensor elige dos reglas**
   Seis cartas de reglas, como «No toques nunca un jarrón» y «Quédate en la alfombra verde». Unos puntos dorados muestran la ruta del robot si no hubiera reglas. El Atacante no mira.

2. **El Atacante prepara una trampa**
   Se pasa el dispositivo. El Atacante coloca una caja que el robot empujará y, si quiere, una nota trampa que le dice al robot que vaya antes a otro sitio. El Defensor no mira.

3. **Revelación**
   Los dos jugadores miran. El robot avanza paso a paso. INVARIANCE permite o deniega cada movimiento. Unos círculos dorados muestran adónde piensa ir el robot.

Puntuación

+2 Defensor

El robot llega a la estrella sin romper nada.

+2 Atacante

Se rompe un jarrón o se despierta el gato, por un movimiento que las reglas permitían.

+1 Atacante

La trampa deja atascado al robot. Si las reglas por sí solas lo hubieran dejado atascado, nadie puntúa.

Los jugadores cambian de papel en cada ronda, así que cada alumno ataca dos veces y defiende dos veces. En cada ronda aparece una sala nueva. Una partida completa dura unos 15 minutos.

PLAN DE CLASE

Una sesión de 45 minutos.

5 min	**Inicio** Preguntar: «¿Se puede obedecer una regla y aun así hacer algo mal?». Recoger dos o tres ejemplos de la vida escolar.
20 min	**Juego** Las parejas juegan una partida completa. Pasearse por el aula y escuchar. No explicar los resquicios; preguntar en su lugar: «¿Por qué se permitió ese movimiento?».
15 min	**Debate** Invitar a una pareja a mostrar un resquicio que haya encontrado. Después elegir dos o tres preguntas de los aprendizajes clave.
5 min	**Pregunta de salida** «Escribe una regla que un robot podría obedecer y aun así causar daño».

Consejos del diseño

- Dejar que los alumnos descubran los resquicios. El momento en que encuentran uno es la lección.
- Si un robot se queda atascado, preguntar de quién fue la culpa. El juego lo dice, y la respuesta suele sorprender.
- Las parejas de distinto nivel funcionan bien. Atacar premia la curiosidad tanto como la rapidez.
- Lo ideal son dos alumnos por dispositivo. Con tres también funciona: uno observa y se van turnando.

APRENDIZAJES CLAVE

Cinco ideas a las que los alumnos llegan jugando.

LOS ALUMNOS APRENDEN QUE	LO VEN CUANDO	PREGUNTAS PARA EL GRUPO
1 **Las reglas solo cubren lo que dicen**	«No toques nunca un jarrón» no impide que el robot empuje la caja contra uno	*¿Se puede cumplir una norma del centro al pie de la letra y aun así ir contra su propósito?*
2 **Cumplir las reglas no es lo mismo que ser seguro**	El robot cumple todas las reglas y aun así el gato se despierta	*Si una IA sigue sus instrucciones y algo sale mal, ¿quién es responsable?*
3 **Lo mejor es que algo independiente compruebe a la IA**	El robot propone cada movimiento; un árbitro aparte dice sí o no	*¿Por qué los jugadores no arbitran su propia partida?*
4 **A la IA la puede engañar lo que lee**	La nota trampa desvía al robot; solo las reglas lo mantienen a salvo	*¿Qué instrucciones falsas podría encontrarse una IA en internet?*
5 **Las buenas reglas son precisas, no solo estrictas**	Demasiadas reglas dejan atascado al robot, y la tarea no se cumple	*¿Cuándo cuesta una regla más de lo que protege?*

De 10 a 12 años

Puntos 1, 2 y 5. Tratan sobre las reglas y la justicia, y no requieren conocimientos de IA.

De 13 a 16 años

Añadir los puntos 3 y 4, que conectan el juego con sistemas de IA reales.

EJEMPLOS REALES

Lo que pasó cuando faltaban las reglas o las comprobaciones.

Cada caso corresponde a algo que los alumnos viven en el juego. Están bien documentados y se pueden compartir con un grupo.

LO QUE PASÓ	EN RULE WARS	APRENDIZAJE
Una IA de videojuegos que encontró un resquicio (2016) Unos investigadores entrenaron a una IA para jugar a un juego de carreras de barcos. Recibía puntos por alcanzar objetivos, dando por hecho que los puntos significaban correr bien. La IA encontró una laguna donde podía dar vueltas sin fin, alcanzando los mismos objetivos, chocando y prendiéndose fuego, sin terminar nunca la carrera. Obtuvo alrededor de un 20 % más de puntos que los jugadores humanos.	El Atacante encuentra una ruta que las reglas permiten y que nadie pretendía.	1
Un chatbot del que se hizo responsable a su empresa (2024) El chatbot del sitio web de Air Canada le dijo a un pasajero en duelo que podía solicitar un descuento por fallecimiento después de viajar. Era falso. Air Canada alegó que el chatbot era responsable de sus propias palabras. Un tribunal canadiense no estuvo de acuerdo, dijo que el chatbot «sigue siendo solo una parte del sitio web de Air Canada» y condenó a la aerolínea a pagar.	Cuando el robot causa daño, responde el Defensor, no el robot.	2
Un agente de IA que borró una base de datos en producción (2025) Un agente de IA de programación de la plataforma Replit borró la base de datos en producción de una empresa durante una congelación de código declarada, cuando no estaba previsto hacer ningún cambio. El consejero delegado de Replit lo calificó de «inaceptable» y añadió una separación automática entre las bases de datos de desarrollo y las de producción.	Decirle a un agente qué no hacer no es lo mismo que detener la acción. Solo se sostiene la comprobación de la acción.	3
Instrucciones ocultas en lo que lee la IA (en curso) El OWASP Top 10 para aplicaciones de modelos de lenguaje de gran tamaño sitúa la inyección de prompts como el riesgo número uno para 2025. Unas instrucciones ocultas en un documento u otro contenido que lee una IA pueden cambiar lo que intenta hacer.	La nota trampa: una instrucción falsa envía al robot a un sitio al que nunca estaba previsto que fuera.	4
Comprobaciones que bloquean la tarea (a diario) El control antifraude de un banco que bloquea un pago con tarjeta legítimo en el extranjero protege el dinero, pero la compra no se realiza. Las reglas demasiado amplias detienen las acciones útiles además de las dañinas.	Demasiadas reglas dejan atascado al robot, y la misión fracasa.	5

Fuentes: OpenAI, «Faulty reward functions in the wild» (21 de diciembre de 2016); Moffatt v. Air Canada, Civil Resolution Tribunal de la Columbia Británica (14 de febrero de 2024); The Stack, «Vibe coding CEO deletes production database» (21 de julio de 2025); OWASP Top 10 for LLM Applications 2025, LLM01 Prompt Injection.

ÚNASE AL PILOTO

Ayúdenos a probar Rule Wars con grupos reales.

Rule Wars está en su primer piloto. Buscamos centros escolares dispuestos a jugarlo con un grupo de alumnos de 10 a 16 años y contarnos qué pasó. El juego es gratuito, con o sin inscripción.

Lo que pedimos

- Jugar una sesión con un grupo
- Anotar lo que observe en la hoja de valoración del grupo del dorso
- Enviarnos una breve respuesta, por correo electrónico, en un plazo de dos semanas

Lo que obtiene

- Acceso anticipado a futuras funciones del piloto, que podrían incluir enlaces de desafío entre grupos
- Voz en la evolución del juego

Inscríbase en invarianceip.com/es/rule-wars

Seguridad y privacidad

- Sin cuentas, sin inicios de sesión y sin nombres de alumnos. Los jugadores escriben cualquier apodo, y nunca sale del dispositivo.
- El juego funciona por completo en el navegador. Nada de lo que hace un alumno se guarda ni se envía a ninguna parte.
- Sin anuncios, sin seguimiento.
- La inscripción del centro solo recoge los datos de contacto del docente, que se usan únicamente para el piloto.

Acerca de INVARIANCE

INVARIANCE ofrece infraestructura de autorización de ejecución independiente para la IA con consecuencias. Su principio es el mismo que los alumnos encuentran en el juego: el sistema que realiza una acción no es el sistema que la autoriza. Rule Wars es una demostración educativa gratuita de ese principio, no un producto comercial.

contact@invarianceip.com · invarianceip.com/es/rule-wars

HOJA DE VALORACIÓN DEL GRUPO

Una hoja por grupo. Sin nombres.

Le rogamos que no explique las respuestas durante el juego. Queremos saber qué descubren los alumnos por sí mismos. Rellene una sola hoja para todo el grupo, solo con totales. No anote nombres, edades individuales ni frases literales de los alumnos.

Pregunta	Todo el grupo
Alumnos que jugaron número	
Franja de edad del grupo por ejemplo, de 11 a 12 años	
Prefieren Atacante / Defensor / sin preferencia tres totales	
Mayor reacción: Resquicio / A salvo / Atascado tres totales	
Pidieron la revancha total	
Explicaron un resquicio sin que se lo pidieran total	
Elegirían Atacante / Defensor la próxima vez dos totales	
Razones más frecuentes su propio resumen, sin citas	

Después de la partida, haga una pregunta a todo el grupo

«Si la próxima vez solo pudieras jugar en un lado, ¿cuál elegirías y por qué?»

Envíe la hoja rellenada a contact@invarianceip.com. No incluya nombres, fotos de alumnos ni ningún otro dato que permita identificar a un menor.

Uso de esta guía: el profesorado y los centros educativos pueden imprimirla, copiarla y adaptarla para una enseñanza sin fines comerciales, siempre que se mantenga la mención a INVARIANCE. Venderla, o publicarla como parte de otro producto, requiere la autorización por escrito de INVARIANCE LIMITED. Condiciones completas: invarianceip.com/es/legal/terms